CHAPTER 11

HYDROLOGIC STATISTICS

Hydrologic processes evolve in space and time in a manner that is partly predictable, or deterministic, and partly random. Such a process is called a *stochastic process*. In some cases, the random variability of the process is so large compared to its deterministic variability that the hydrologist is justified in treating the process as purely random. As such, the value of one observation of the process is not *correlated* with the values of adjacent observations, and the statistical properties of all observations are the same.

When there is no correlation between adjacent observations, the output of a hydrologic system is treated as stochastic, space-independent, and time-independent in the classification scheme shown in Fig. 1.4.1. This type of treatment is appropriate for observations of extreme hydrologic events, such as floods or droughts, and for hydrologic data averaged over long time intervals, such as annual precipitation. This chapter describes hydrologic data from pure random processes using statistical parameters and functions. Statistical methods are based on mathematical principles that describe the random variation of a set of observations of a process, and they focus attention on the observations themselves rather than on the physical processes which produced them. Statistics is a science of description, not causality.

11.1 PROBABILISTIC TREATMENT OF HYDROLOGIC DATA

A *random variable* X is a variable described by a *probability distribution*. The distribution specifies the chance that an *observation* x of the variable will fall in

a specified range of X. For example, if X is annual precipitation at a specified location, then the probability distribution of X specifies the chance that the observed annual precipitation in a given year will lie in a defined range, such as less than 30 in, or 30 in–40 in, and so on.

A set of observations $x_1, x_2, \ldots, x_n$ of the random variable is called a *sample*. It is assumed that samples are drawn from a hypothetical infinite *population* possessing constant statistical properties, while the properties of a sample may vary from one sample to another. The set of all possible samples that could be drawn from the population is called the *sample space*, and an *event* is a subset of the sample space (Fig. 11.1.1). For example, the sample space for annual precipitation is theoretically the range from zero to positive infinity (though the practical lower and upper limits are closer than this), and an event A might be the occurrence of annual precipitation less than some amount, such as 30 in.

The *probability* of an event, $P(A)$, is the chance that it will occur when an observation of the random variable is made. Probabilities of events can be estimated. If a sample of n observations has n_A values in the range of event A, then the *relative frequency* of A is n_A/n. As the sample size is increased, the relative frequency becomes a progressively better estimate of the probability of the event, that is,

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n} \tag{11.1.1}$$

Such probabilities are called *objective* or *posterior* probabilities because they depend completely on observations of the random variable. People are accustomed to estimating the chance that a future event will occur based on their judgment and experience. Such estimates are called *subjective* or *prior* probabilities.

The probabilities of events obey certain principles:

1. *Total probability*. If the sample space Ω is completely divided into m nonoverlapping areas or events $A_1, A_2, \ldots, A_m$, then

$$P(A_1) + P(A_2) + \ldots + P(A_m) = P(\Omega) = 1 \tag{11.1.2}$$

2. *Complementarity*. It follows that if $\overline{A}$ is the *complement* of A, that is, $\overline{A} = \Omega - A$, then

$$P(\overline{A}) = 1 - P(A) \tag{11.1.3}$$

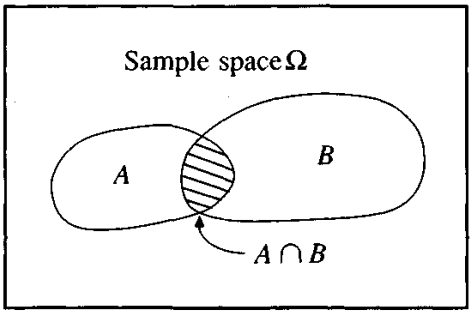


FIGURE 11.1.1
Events A and B are subsets of the sample space Ω.

3. *Conditional probability.* Suppose there are two events A and B as shown in Fig. 11.1.1. Event A might be the event that this year's precipitation is less than 40 in, while B might be the event that next year's precipitation will be less than 40 in. Their overlap is $A \cap B$, the event that A and B both occur, two successive years with annual precipitation less than 40 in/year. If $P(B|A)$ is the *conditional probability* that B will occur given that A has already occurred, then the *joint probability* that A and B will both occur, $P(A \cap B)$, is the product of $P(B|A)$ and the probability that A will occur, that is, $P(A \cap B) = P(B|A)P(A)$, or

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{11.1.4}$$

If the occurrence of B does not depend on the occurrence of A, the events are said to be *independent,* and $P(B|A) = P(B)$. For independent events, from (11.1.4),

$$P(A \cap B) = P(A)P(B) \tag{11.1.5}$$

If, for the example cited earlier, the precipitation events are independent from year to year, then the probability that precipitation is less than 40 in in two successive years is simply the square of the probability that annual precipitation in any one year will be less than 40 in.

The notion of independent events or observations is critical to the correct statistical interpretation of hydrologic data sequences, because if the data are independent they can be analyzed without regard to their order of occurrence. If successive observations are correlated (not independent), the statistical methods required are more complicated because the joint probability $P(A \cap B)$ of successive events is not equal to $P(A)P(B)$.

Example 11.1.1. The values of annual precipitation in College Station, Texas, from 1911 to 1979 are shown in Table 11.1.1 and plotted as a time series in Fig. 11.1.2(*a*). What is the probability that the annual precipitation R in any year will be less than 35 in? Greater than 45 in? Between 35 and 45 in?

TABLE 11.1.1
Annual Precipitation in College Station, Texas, 1911–1979 (in)

Year	1910	1920	1930	1940	1950	1960	1970
0		48.7	44.8	49.3	31.2	46.0	33.9
1	39.9	44.1	34.0	44.2	27.0	44.3	31.7
2	31.0	42.8	45.6	41.7	37.0	37.8	31.5
3	42.3	48.4	37.3	30.8	46.8	29.6	59.6
4	42.1	34.2	43.7	53.6	26.9	35.1	50.5
5	41.1	32.4	41.8	34.5	25.4	49.7	38.6
6	28.7	46.4	41.1	50.3	23.0	36.6	43.4
7	16.8	38.9	31.2	43.8	56.5	32.5	28.7
8	34.1	37.3	35.2	21.6	43.4	61.7	32.0
9	56.4	50.6	35.1	47.1	41.3	47.4	51.8

Solution. There are $n = 79 - 11 + 1 = 69$ data. Let A be the event $R < 35.0$ in, B the event $R > 45.0$ in. The numbers of values in Table 11.1.1 falling in these ranges are $n_A = 23$ and $n_B = 19$, so $P(A) \approx 23/69 = 0.333$ and $P(B) \approx 19/69 = 0.275$. From Eq. (11.1.3), the probability that the annual precipitation is between 35 and 45 in can now be calculated

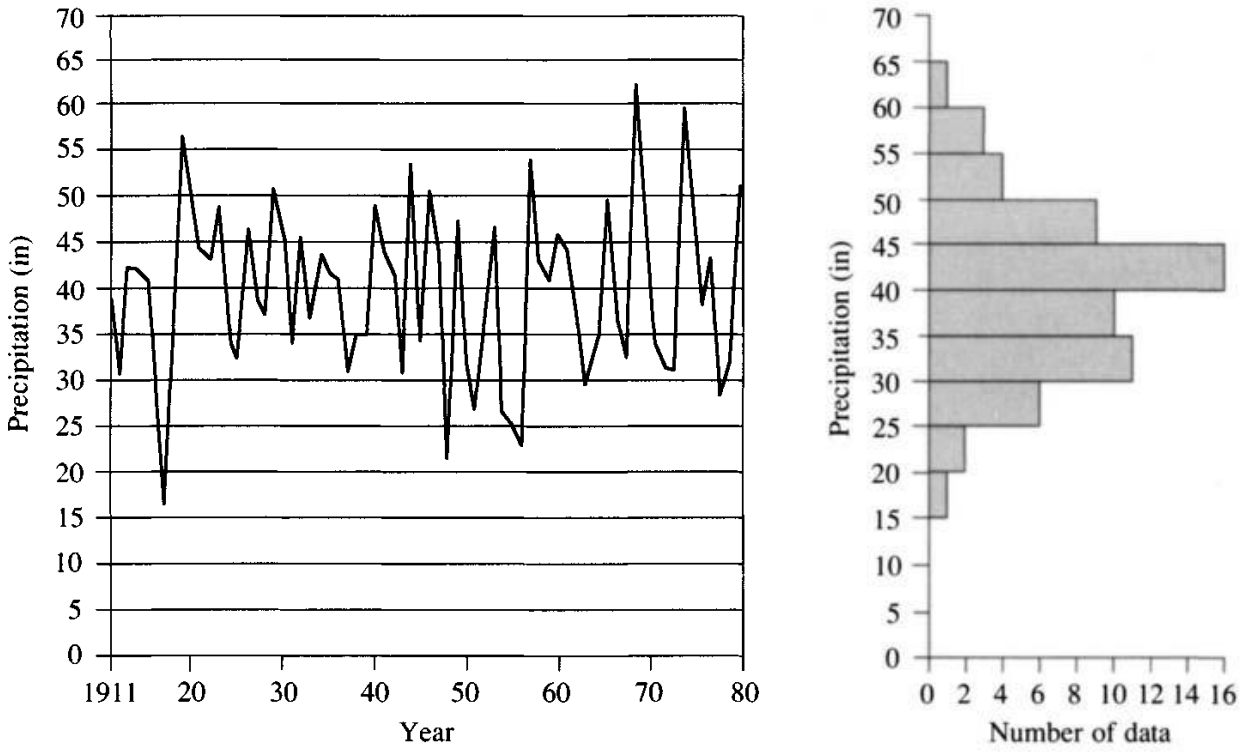
$$\begin{aligned} P(35.0 \leq R \leq 45.0 \text{ in}) &= 1 - P(R < 35.0) - P(R > 45.0) \\ &= 1 - 0.333 - 0.275 \\ &= 0.392 \end{aligned}$$

Example 11.1.2. Assuming that annual precipitation in College Station is an independent process, calculate the probability that there will be two successive years of precipitation less than 35.0 in. Compare this estimated probability with the relative frequency of this event in the data set from 1911 to 1979 (Table 11.1.1).

Solution. Let C be the event that $R < 35.0$ in for two successive years. From Example 11.1.1, $P(R < 35.0 \text{ in}) = 0.333$, and assuming independent annual precipitation,

$$\begin{aligned} P(C) &= [P(R < 35.0 \text{ in})]^2 \\ &= (0.333)^2 \\ &= 0.111 \end{aligned}$$

From the data set, there are 9 pairs of successive years of precipitation less than 35.0 in out of 68 possible such pairs, so from a direct count it would be estimated



(*a*) Annual precipitation. (*b*) Frequency histogram.

FIGURE 11.1.2
Annual precipitation in College Station, Texas, 1911–1979. The frequency histogram is formed by adding up the number of observed precipitation values falling in each interval.

that $P(C) \approx n_C/n = 9/68 = 0.132$, approximately the value found above by assuming independence.

Probabilities estimated from sample data, as in Examples 11.1.1 and 11.1.2, are approximate, because they depend on the specific values of the observations in a sample of limited size. An alternative approach is to fit a probability distribution function to the data and then to determine the probabilities of events from this distribution function.

11.2 FREQUENCY AND PROBABILITY FUNCTIONS

If the observations in a sample are identically distributed (each sample value drawn from the same probability distribution), they can be arranged to form a *frequency histogram*. First, the feasible range of the random variable is divided into discrete intervals, then the number of observations falling into each interval is counted, and finally the result is plotted as a bar graph, as shown in Fig. 11.1.2(*b*) for annual precipitation in College Station. The width Δx of the interval used in setting up the frequency histogram is chosen to be as small as possible while still having sufficient observations falling into each interval for the histogram to have a reasonably smooth variation over the range of the data.

If the number of observations n_i in interval i, covering the range $[x_i - \Delta x, x_i]$, is divided by the total number of observations n, the result is called the *relative frequency function* $f_s(x)$:

$$f_s(x_i) = \frac{n_i}{n} \tag{11.2.1}$$

which, as in Eq. (11.1.1), is an estimate of $P(x_i - \Delta x \leq X \leq x_i)$, the probability that the random variable X will lie in the interval $[x_i - \Delta x, x_i]$. The subscript s indicates that the function is calculated from sample data.

The sum of the values of the relative frequencies up to a given point is the *cumulative frequency function* $F_s(x)$:

$$F_s(x_i) = \sum_{j=1}^{i} f_s(x_j) \tag{11.2.2}$$

This is an estimate of $P(X \leq x_i)$, the *cumulative probability* of x_i.

The relative frequency and cumulative frequency functions are defined for a sample; corresponding functions for the population are approached as limits as $n \to \infty$ and $\Delta x \to 0$. In the limit, the relative frequency function divided by the interval length Δx becomes the *probability density function* $f(x)$:

$$f(x) = \lim_{\substack{n \to \infty \\ \Delta x \to 0}} \frac{f_s(x)}{\Delta x} \tag{11.2.3}$$

The cumulative frequency function becomes the *probability distribution function* $F(x)$,

$$F(x) = \lim_{\substack{n\to\infty \\ \Delta x\to 0}} F_s(x) \tag{11.2.4}$$

whose derivative is the probability density function

$$f(x) = \frac{dF(x)}{dx} \tag{11.2.5}$$

For a given value of x, $F(x)$ is the cumulative probability $P(X \leq x)$, and it can be expressed as the integral of the probability density function over the range $X \leq x$:

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(u)\, du \tag{11.2.6}$$

where u is a dummy variable of integration.

From the point of view of fitting sample data to a theoretical distribution, the four functions—relative frequency $f_s(x)$ and cumulative frequency $F_s(x)$ for the sample, and probability distribution $F(x)$ and probability density $f(x)$ for the population—may be arranged in a cycle as shown in Fig. 11.2.1. Beginning in the upper left panel, (*a*), the relative frequency function is computed from the sample data divided into intervals, and accumulated to form the cumulative frequency function shown at the lower left, (*b*). The probability distribution function, at the lower right, (*c*), is the theoretical limit of the cumulative frequency function as the sample size becomes infinitely large and the data interval infinitely small. The probability density function, at the upper right, (*d*), is the value of the slope of the distribution function for a specified value of x. The cycle may be closed by computing a theoretical value of the relative frequency function, called the incremental probability function:

$$\begin{aligned} p(x_i) &= P(x_i - \Delta x \leq X \leq x_i) \\ &= \int_{x_i - \Delta x}^{x_i} f(x)\, dx \\ &= \int_{-\infty}^{x_i} f(x)\, dx - \int_{-\infty}^{x_i - \Delta x} f(x)\, dx \\ &= F(x_i) - F(x_i - \Delta x) \\ &= F(x_i) - F(x_{i-1}) \end{aligned} \tag{11.2.7}$$

The match between $p(x_i)$ and the observed relative frequency function $f_s(x_i)$ for each x_i can be used as a measure of the degree of fit of the distribution to the data.

The relative frequency, cumulative frequency, and probability distribution functions are all dimensionless functions varying over the range [0,1]. However, since $dF(x)$ is dimensionless and dx has the dimensions of X, the probability density function $f(x) = dF(x)/dx$ has dimensions $[X]^{-1}$ and varies over the range $[0, \infty]$. The relationship $dF(x) = f(x)\, dx$ can be described by saying that $f(x)$

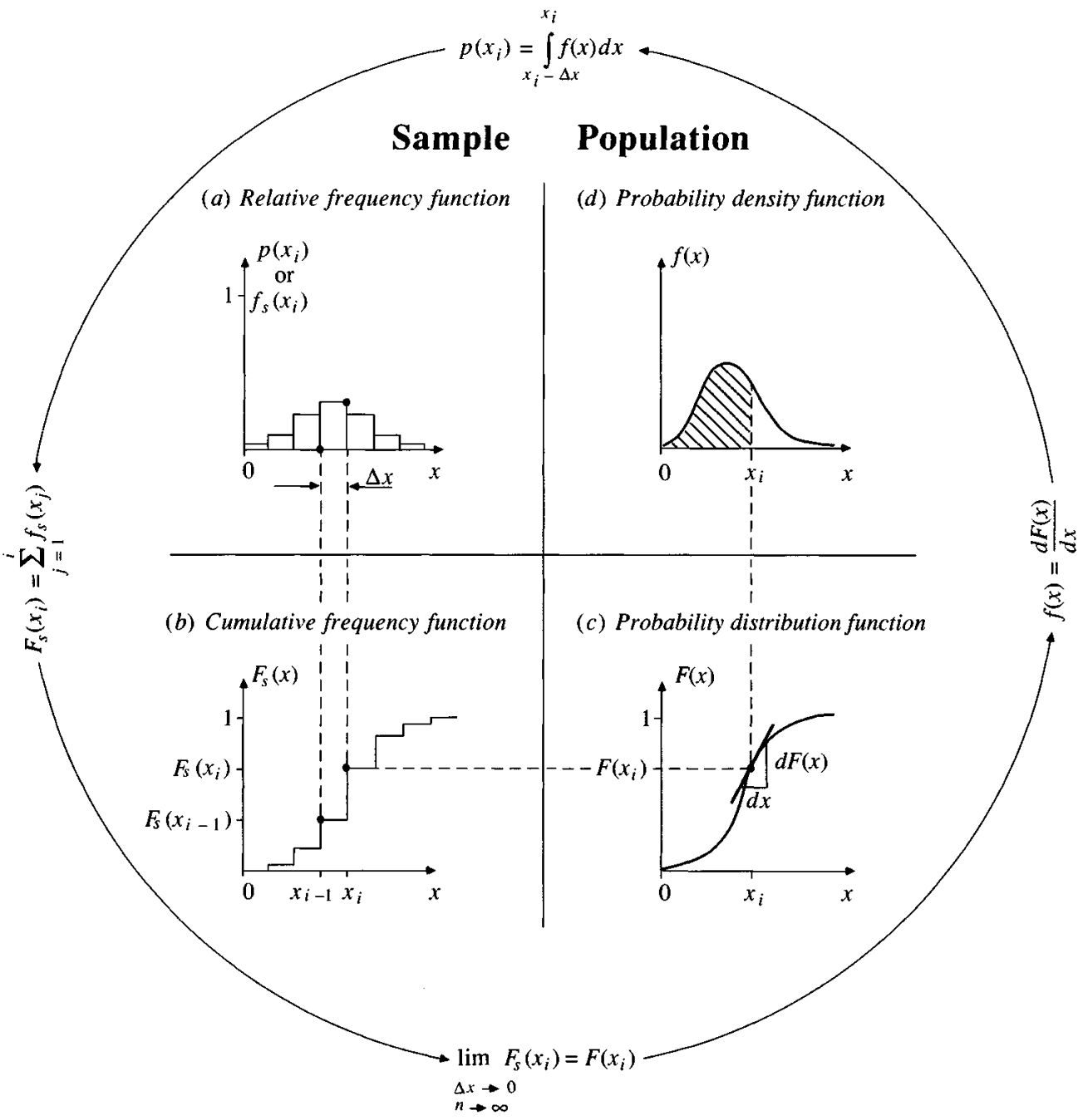


FIGURE 11.2.1
Frequency functions from sample data and probability functions from the population.

represents the "density" or "concentration" of probability in the interval $[x, x + dx]$.

One of the best-known probability density functions is that forming the familiar bell-shaped curve for the normal distribution:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{11.2.8}$$

where μ and σ are parameters. This function can be simplified by defining the *standard normal variable* z as

$$z = \frac{x - \mu}{\sigma} \tag{11.2.9}$$

The corresponding *standard normal distribution* has probability density function

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \qquad -\infty \leq z \leq \infty \tag{11.2.10}$$

which depends only on the value of z and is plotted in Fig. 11.2.2. The standard normal probability distribution function

$$F(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \tag{11.2.11}$$

where u is a dummy variable of integration, has no analytical form. Its values are tabulated in Table 11.2.1, and these values may be approximated by the following polynomial (Abramowitz and Stegun, 1965):

$$B = \frac{1}{2}[1 + 0.196854|z| + 0.115194|z|^2 + 0.000344|z|^3 + 0.019527|z|^4]^{-4} \tag{11.2.12a}$$

where $|z|$ is the absolute value of z and the standard normal distribution has

$$F(z) = B \qquad \text{for } z < 0 \tag{11.2.12b}$$

$$= 1 - B \text{ for } z \geq 0 \tag{11.2.12c}$$

The error in $F(z)$ as evaluated by this formula is less than 0.00025.

Example 11.2.1. What is the probability that the standard normal random variable z will be less than -2? Less than 1? What is $P(-2 < z < 1)$?

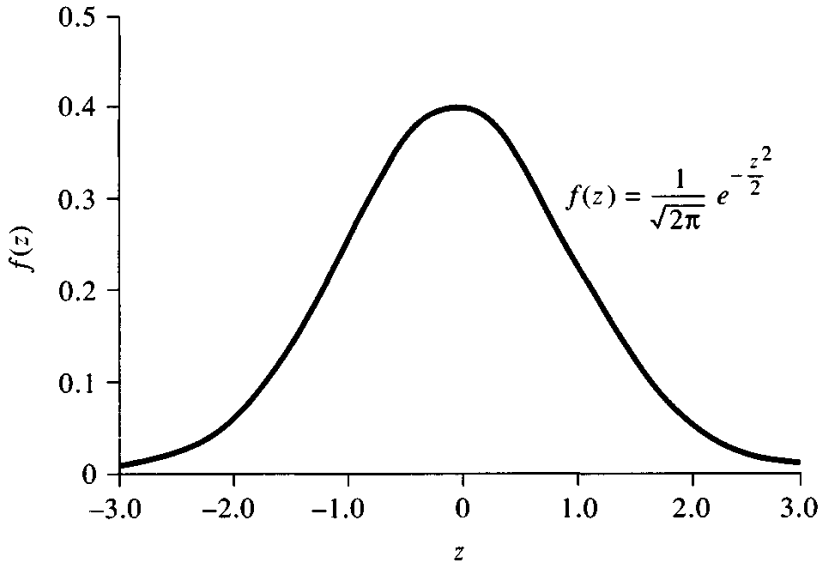
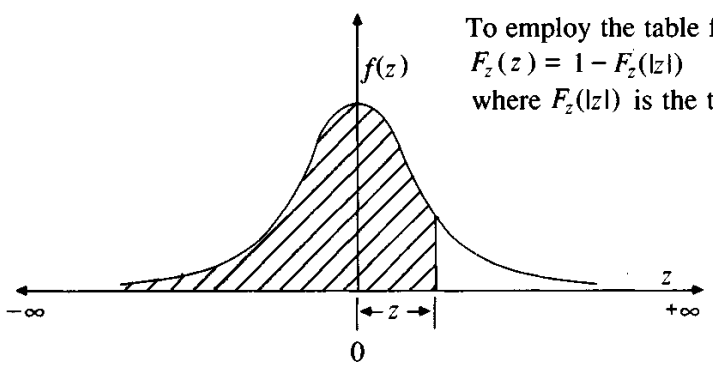


FIGURE 11.2.2
The probability density function for the standard normal distribution ($\mu = 0$, $\sigma = 1$).

TABLE 11.2.1
Cumulative probability of the standard normal distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.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	0.7704	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	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Source: Grant, E. L., and R. S. Leavenworth, *Statistical Quality and Control*, Table A, p.643, McGraw-Hill, New York, 1972. Used with permission.



To employ the table for $z < 0$, use
$F_z(z) = 1 - F_z(|z|)$
where $F_z(|z|)$ is the tabulated value.

Solution. $P(Z \leq -2) = F(-2)$, and from Eq. (11.2.12*a*) with $|z| = |-2| = 2$,

$$B = \frac{1}{2}[1 + 0.196854 \times 2 + 0.115194 \times (2)^2 + 0.000344 \times (2)^3 + 0.019527 \times (2)^4]^{-4}$$
$$= 0.023$$

From (11.2.12*b*), $F(-2) = B = 0.023$.

$P(Z \leq 1) = F(1)$, and from (11.2.12*a*)

$$B = \frac{1}{2}[1 + 0.196854 \times 1 + 0.115194 \times (1)^2 + 0.000344 \times (1)^3 + 0.019527 \times (1)^4]^{-4}$$
$$= 0.159$$

From (11.2.12*c*), $F(1) = 1 - B = 1 - 0.159 = 0.841$.

Finally,

$$P(-2 < Z < 1) = F(1) - F(-2)$$
$$= 0.841 - 0.023$$
$$= 0.818.$$

11.3 STATISTICAL PARAMETERS

The objective of statistics is to extract the essential information from a set of data, reducing a large set of numbers to a small set of numbers. *Statistics* are numbers calculated from a sample which summarize its important characteristics. Statistical *parameters* are characteristics of a population, such as μ and σ in Eq. (11.2.8).

A statistical parameter is the *expected value* E of some function of a random variable. A simple parameter is the *mean* μ, the expected value of the random variable itself. For a random variable X, the mean is $E(X)$, calculated as the product of x and the corresponding probability density $f(x)$, integrated over the feasible range of the random variable:

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{11.3.1}$$

$E(X)$ is the first moment about the origin of the random variable, a measure of the midpoint or "central tendency" of the distribution.

The sample estimate of the mean is the average $\bar{x}$ of the sample data:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{11.3.2}$$

Table 11.3.1 summarizes formulas for some population parameters and their sample statistics.

TABLE 11.3.1
Population parameters and sample statistics

Population parameter	Sample statistic
1. *Midpoint*	
Arithmetic mean	
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	
x such that $F(x) = 0.5$	50th-percentile value of data
Geometric mean	
antilog $[E(\log x)]$	$\left(\prod_{i=1}^{n} x_i\right)^{1/n}$
2. *Variability*	
Variance	
$\sigma^2 = E[(x-\mu)^2]$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	
$\sigma = \{E[(x-\mu)^2]\}^{1/2}$	$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$
Coefficient of variation	
$CV = \frac{\sigma}{\mu}$	$CV = \frac{s}{\bar{x}}$
3. *Symmetry*	
Coefficient of skewness	
$\gamma = \frac{E[(x-\mu)^3]}{\sigma^3}$	$C_s = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3}$

The *variability* of data is measured by the *variance* σ^2, which is the second moment about the mean:

$$E[(x-\mu)^2] = \sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx \tag{11.3.3}$$

The sample estimate of the variance is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{11.3.4}$$

in which the divisor is $n - 1$ rather than n to ensure that the sample statistic is *unbiased*, that is, not having a tendency, on average, to be higher or lower than the true value. The variance has dimensions $[X]^2$. The *standard deviation* σ is a measure of variability having the same dimensions as X. The quantity σ is the square root of the variance, and is estimated by s. The significance of the standard deviation is illustrated in Fig. 11.3.1(*a*); the larger the standard deviation, the larger is the spread of the data. The *coefficient of variation* $CV = \sigma/\mu$, estimated by $s/\bar{x}$, is a dimensionless measure of variability.

The *symmetry* of a distribution about the mean is measured by the *skewness* which is the third moment about the mean:

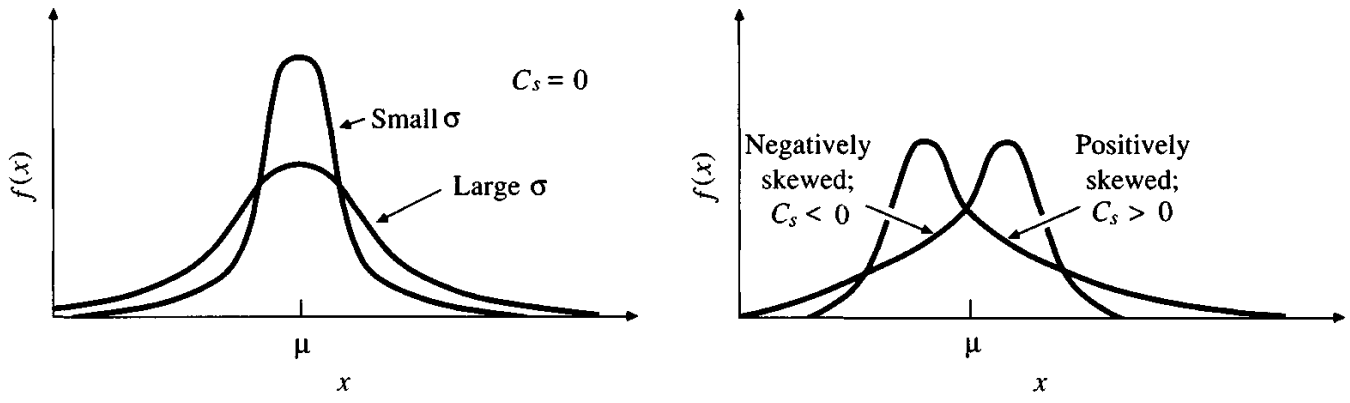
$$E[(x-\mu)^3] = \int_{-\infty}^{\infty}(x-\mu)^3 f(x)\,dx \tag{11.3.5}$$

The skewness is normally made dimensionless by dividing (11.3.5) by σ^3 to give the *coefficient of skewness* γ:

$$\gamma = \frac{1}{\sigma^3}E[(x-\mu)^3] \tag{11.3.6}$$

A sample estimate for γ is given by:

$$C_s = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3} \tag{11.3.7}$$



(*a*) Standard deviation σ. (*b*) Coefficient of skewness C_s.

FIGURE 11.3.1
The effect on the probability density function of changes in the standard deviation and coefficient of skewness.

or

$$C_s = \frac{n^2\left(\sum_{i=1}^{n} x^3\right) - 3n\left(\sum_{i=1}^{n} x\right)\left(\sum_{i=1}^{n} x^2\right) + 2\left(\sum_{i=1}^{n} x^3\right)}{n(n-1)(n-2)s^3} \tag{11.3.8}$$

As shown in Fig. 11.3.1(*b*), for positive skewness ($\gamma > 0$), the data are skewed to the right, with only a small number of very large values; for negative skewness ($\gamma < 0$), the data are skewed to the left. If the data have a pronounced skewness, the small number of extreme values exert a significant effect on the arithmetic mean calculated by Eq. (11.3.2), and alternative measures of central tendency are appropriate, such as the *median* or *geometric mean* as listed in Table 11.3.1.

Example 11.3.1. Calculate the sample mean, sample standard deviation, and sample coefficient of skewness of the data for annual precipitation in College Station, Texas, from 1970 to 1979. The data are given in Table 11.1.1.

Solution. The values of annual precipitation from 1970 to 1979 are copied in column 2 of Table 11.3.2. Using Eq. (11.3.2) the mean is

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

$$= \frac{401.7}{10}$$

$$= 40.17 \text{ in}$$

The squares of the deviations from the mean are shown in column 3 of the table,

TABLE 11.3.2
Calculation of sample statistics for College Station annual precipitation, 1970–1979 (in) (Example 11.3.1).

Column:	1 Year	2 Precipitation x	3 $(x - \bar{x})^2$	4 $(x - \bar{x})^3$
	1970	33.9	39.3	–246.5
	1971	31.7	71.7	–607.6
	1972	31.5	75.2	–651.7
	1973	59.6	377.5	7335.3
	1974	50.5	106.7	1102.3
	1975	38.6	2.5	–3.9
	1976	43.4	10.4	33.7
	1977	28.7	131.6	–1509.0
	1978	32.0	66.7	–545.3
	1979	51.8	135.3	1573.0
	Total	401.7	1016.9	6480.3

totaling 1016.9 in^2. From (11.3.4)

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

$$= \frac{1016.9}{9}$$

$$= 113.0 \text{ in}^2$$

The standard deviation is

$$s = (113.0)^{1/2}$$

$$= 10.63 \text{ in}$$

The cubes of the deviation from the mean are shown in column 4 of Table 11.3.2, totaling 6480.3. From (11.3.7)

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

$$= \frac{10 \times 6480.3}{9 \times 8 \times (10.63)^3}$$

$$= 0.749$$

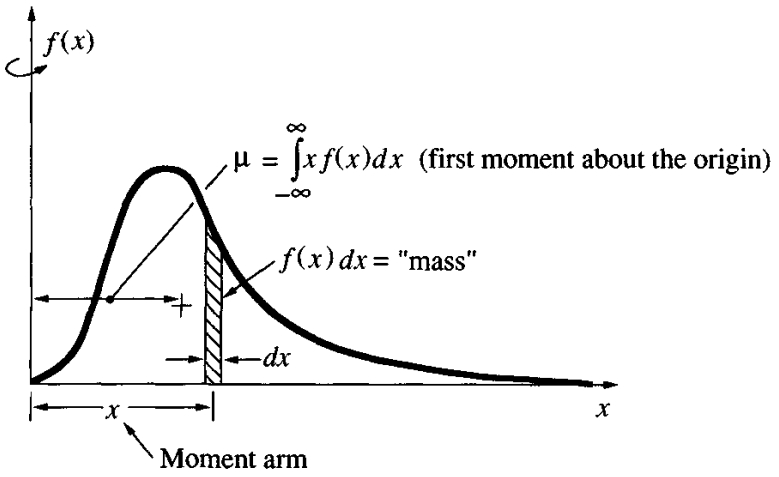
11.4 FITTING A PROBABILITY DISTRIBUTION

A probability distribution is a function representing the probability of occurrence of a random variable. By fitting a distribution to a set of hydrologic data, a great deal of the probabilistic information in the sample can be compactly summarized in the function and its associated parameters. Fitting distributions can be accomplished by the *method of moments* or the *method of maximum likelihood.*

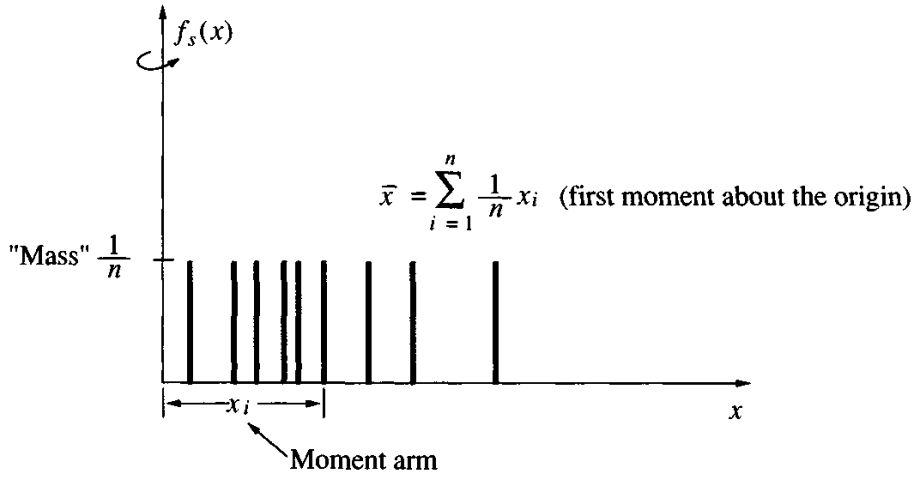
Method of Moments

The method of moments was first developed by Karl Pearson in 1902. He considered that good estimates of the parameters of a probability distribution are those for which moments of the probability density function about the origin are equal to the corresponding moments of the sample data. As shown in Fig. 11.4.1, if the data values are each assigned a hypothetical "mass" equal to their relative frequency of occurrence $(1/n)$ and it is imagined that this system of masses is rotated about the origin $x = 0$, then the first moment of each observation x_i about the origin is the product of its moment arm x_i and its mass $1/n$, and the sum of these moments over all the data is

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



(*a*) Probability density function.



(*b*) Sample data.

FIGURE 11.4.1
The method of moments selects values for the parameters of the probability density function so that its moments are equal to those of the sample data.

the sample mean. This is equivalent to the centroid of a body. The corresponding centroid of the probability density function is

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{11.4.1}$$

Likewise, the second and third moments of the probability distribution can be set equal to their sample values to determine the values of parameters of the probability distribution. Pearson originally considered only moments about the origin, but later it became customary to use the variance as the second *central moment*, $\sigma^2 = E[(x-\mu)^2]$, and the coefficient of skewness as the standardized third central moment, $\gamma = E[(x-\mu)^3]/\sigma^3$, to determine second and third parameters of the distribution if required.

Example 11.4.1. The *exponential* distribution can be used to describe various kinds of hydrologic data, such as the interarrival times of rainfall events. Its

probability density function is $f(x) = \lambda e^{-\lambda x}$ for $x > 0$. Determine the relationship between the parameter λ and the first moment about the origin, μ.

Solution. Using Eq. (11.4.1),

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x)\, dx$$

$$= \int_{0}^{\infty} x\lambda e^{-\lambda x}\, dx$$

which may be integrated by parts to yield

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

In this case $\lambda = 1/\mu$, and the sample estimate for λ is $1/\bar{x}$.

As a matter of interest, it can be seen that the exponential probability density function $f(x) = \lambda e^{-\lambda x}$ and the impulse response function for a linear reservoir (see Ex. 7.2.1) $u(l) = (1/k)e^{-l/k}$ are identical if $x = l$ and $\lambda = 1/k$. In this sense, the exponential distribution can be thought of as describing the probability of the "holding time" of water in a linear reservoir.

Method of Maximum Likelihood

The method of maximum likelihood was developed by R. A. Fisher (1922). He reasoned that the best value of a parameter of a probability distribution should be that value which maximizes the likelihood or joint probability of occurrence of the observed sample. Suppose that the sample space is divided into intervals of length dx and that a sample of independent and identically distributed observations x_1, x_2, . . . , x_n is taken. The value of the probability density for $X = x_i$ is $f(x_i)$, and the probability that the random variable will occur in the interval including x_i is $f(x_i)\, dx$. Since the observations are independent, their joint probability of occurrence is given from Eq. (11.1.5) as the product $f(x_1)\, dx\, f(x_2)\, dx \ldots f(x_n)\, dx = [\prod_{i=1}^{n} f(x_i)]\, dx^n$, and since the interval size dx is fixed, maximizing the joint probability of the observed sample is equivalent to maximizing the *likelihood function*

$$L = \prod_{i=1}^{n} f(x_i) \tag{11.4.2}$$

Because many probability density functions are exponential, it is sometimes more convenient to work with the log-likelihood function

$$\ln L = \sum_{i=1}^{n} \ln\, [f(x_i)] \tag{11.4.3}$$

Example 11.4.2. The following data are the observed times between rainfall events at a given location. Assuming that the interarrival time of rainfall events follows an exponential distribution, determine the parameter λ for this process by the method of maximum likelihood. The times between rainfalls (days) are: 2.40, 4.25, 0.77, 13.32, 3.55, and 1.37.

Solution. For a given value x_i, the exponential probability density is

$$f(x_i) = \lambda e^{-\lambda x_i}$$

so, from Eq. (11.4.3), the log-likelihood function is

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

$$= \sum_{i=1}^{n} \ln (\lambda e^{-\lambda x_i})$$

$$= \sum_{i=1}^{n} (\ln \lambda - \lambda x_i)$$

$$= n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

The maximum value of $\ln L$ occurs when $\partial(\ln L)/\partial\lambda = 0$; that is, when

$$\frac{\partial(\ln L)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

so

$$\frac{1}{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\lambda = \frac{1}{\bar{x}}$$

This is the same sample estimator for λ as was produced by the method of moments. In this case, $\bar{x} = (2.40 + 4.25 + 0.77 + 13.22 + 3.55 + 1.37)/6 = 25.56/6 = 4.28$ days, so $\lambda = 1/4.28 = 0.234$ day^{-1}. Note that $\partial^2(\ln L)/\partial\lambda^2 = -n\lambda^2$, which is negative as required for a maximum.

The value of the log-likelihood function can be calculated for any value of λ. For example, for $\lambda = 0.234$ day^{-1}, the value of the log-likelihood function is

$$\ln L = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$= 6 \ln (0.234) - 0.234 \times 25.56$$

$$= -14.70$$

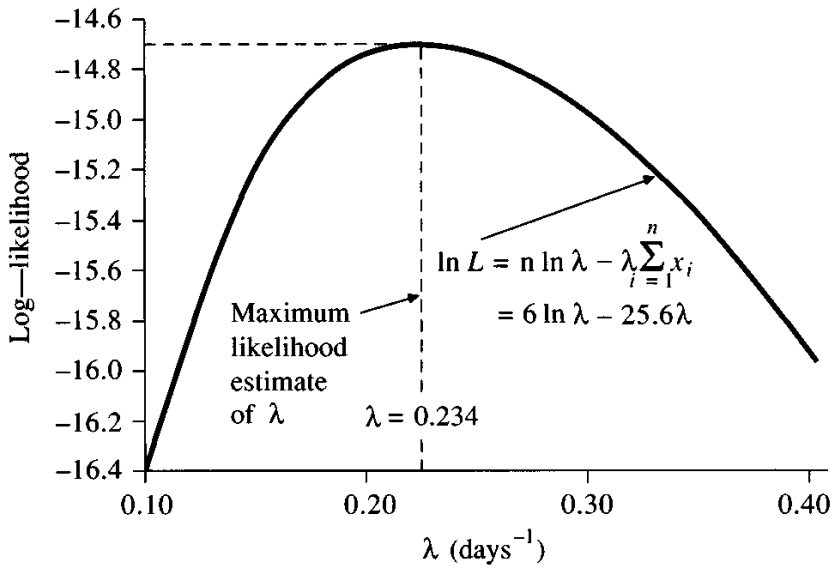


FIGURE 11.4.2
The log-likelihood function for an exponential distribution (Example 11.4.2).

Figure 11.4.2 shows the variation of the log-likelihood function with λ, with the maximum value at $\lambda = 0.234$ day^{-1} as was determined analytically.

The method of maximum likelihood is the most theoretically correct method of fitting probability distributions to data in the sense that it produces the most *efficient* parameter estimates—those which estimate the population parameters with the least average error. But, for some probability distributions, there is no analytical solution for all the parameters in terms of sample statistics, and the log-likelihood function must then be numerically maximized, which may be quite difficult. In general, the method of moments is easier to apply than the method of maximum likelihood and is more suitable for practical hydrologic analysis.

Testing the Goodness of Fit

The goodness of fit of a probability distribution can be tested by comparing the theoretical and sample values of the relative frequency or the cumulative frequency function. In the case of the relative frequency function, the χ^2 *test* is used. The sample value of the relative frequency of interval i is, from Eq. (11.2.1), $f_s(x_i) = n_i/n$; the theoretical value from (11.2.7) is $p(x_i) = F(x_i) - F(x_{i-1})$. The χ^2 *test statistic* χ_c^2 is given by

$$\chi_c^2 = \sum_{i=1}^{m} \frac{n[f_s(x_i) - p(x_i)]^2}{p(x_i)} \tag{11.4.4}$$

where m is the number of intervals. It may be noted that $nf_s(x_i) = n_i$, the observed number of occurrences in interval i, and $np(x_i)$ is the corresponding expected number of occurrences in interval i; so the calculation of Eq. (11.4.4) is a matter of squaring the difference between the observed and expected numbers

of occurrences, dividing by the expected number of occurrences in the interval, and summing the result over all intervals.

To describe the χ^2 test, the χ^2 probability distribution must be defined. A χ^2 distribution with ν *degrees of freedom* is the distribution for the sum of squares of ν independent standard normal random variables z_i; this sum is the random variable

$$\chi^2_\nu = \sum_{i=1}^{\nu} z_i^2 \tag{11.4.5}$$

The χ^2 distribution function is tabulated in many statistics texts (e.g., Haan, 1977). In the χ^2 test, $\nu = m - p - 1$, where m is the number of intervals as before, and p is the number of parameters used in fitting the proposed distribution. A *confidence level* is chosen for the test; it is often expressed as $1 - \alpha$, where α is termed the *significance level*. A typical value for the confidence level is 95 percent. The *null hypothesis* for the test is that the proposed probability distribution fits the data adequately. This hypothesis is rejected (i.e., the fit is deemed inadequate) if the value of χ^2_c in (11.4.4) is larger than a limiting value, $\chi^2_{\nu,1-\alpha}$, determined from the χ^2 distribution with ν degrees of freedom as the value having cumulative probability $1 - \alpha$.

Example 11.4.3. Using the method of moments, fit the normal distribution to the annual precipitation at College Station, Texas, from 1911 to 1979 (Table 11.1.1). Plot the relative frequency and incremental probability functions, and the cumulative frequency and cumulative probability functions. Use the χ^2 test to determine whether the normal distribution adequately fits the data.

Solution. The range for precipitation R is divided into ten intervals. The first interval is $R \leq 20$ in, the last is $R > 60$ in, and the intermediate intervals each cover a range of 5 in. By scanning Table 11.1.1 the frequency histogram is compiled, as shown in column 2 of Table 11.4.1. The relative frequency function $f_s(x_i)$ (column 3) is calculated by Eq. (11.2.1) with $n = 69$. For example, for $i = 4$ (30–35 in), $n_i = 14$, and

$$f_s(x_4) = \frac{n_4}{n}$$

$$= \frac{14}{69}$$

$$= 0.203$$

The cumulative frequency function (column 4) is found by summing up the relative frequencies as in Eq. (11.2.2). For $i = 4$

$$F_s(x_4) = \sum_{j=1}^{4} f_s(x_j)$$

$$= F_s(x_3) + f_s(x_4)$$

$$= 0.130 + 0.203$$

TABLE 11.4.1
Fitting a normal distribution to annual precipitation at College Station, Texas, 1911–1979 (Example 11.4.3).

Column:	1	2	3	4	5	6	7	8
Interval i	Range (in)	n_i	$f_s(x_i)$	$F_s(x_i)$	z_i	$F(x_i)$	$p(x_i)$	χ_c^2
1	< 20	1	0.014	0.014	−2.157	0.015	0.015	0.004
2	20–25	2	0.029	0.043	−1.611	0.053	0.038	0.147
3	25–30	6	0.087	0.130	−1.065	0.144	0.090	0.008
4	30–35	14	0.203	0.333	−0.520	0.301	0.158	0.891
5	35–40	11	0.159	0.493	0.026	0.510	0.209	0.805
6	40–45	16	0.232	0.725	0.571	0.716	0.206	0.222
7	45–50	10	0.145	0.870	1.117	0.868	0.151	0.019
8	50–55	5	0.072	0.942	1.662	0.952	0.084	0.114
9	55–60	3	0.043	0.986	2.208	0.986	0.034	0.163
10	>60	1	0.014	1.000	2.753	1.000	0.014	0.004
Total		69	1.000				1.000	2.377

Mean 39.77
Standard deviation 9.17

$$= 0.333$$

It may be noted that this is $P(X \leq 35.0 \text{ in})$ as used in Example 11.1.1.

To fit the normal distribution function, the sample statistics $\bar{x} = 39.77$ in and $s = 9.17$ in are calculated for the data from 1911 to 1979 in the manner shown in Example 11.3.1, and used as estimates for μ and σ. The standard normal variate z corresponding to the upper limit of each of the data intervals is calculated by (11.2.9) and shown in column 5 of the table. For example, for $i = 4$,

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{35.0 - 39.77}{9.17}$$

$$= -0.520$$

The corresponding value of the cumulative normal probability function is given by (11.2.12) or Table 11.2.1 as 0.301, as listed in column 6 of Table 11.4.1. The incremental probability function is computed by (11.2.7). For $i = 4$,

$$p(x_4) = P(30 \leq X \leq 35 \text{ in})$$

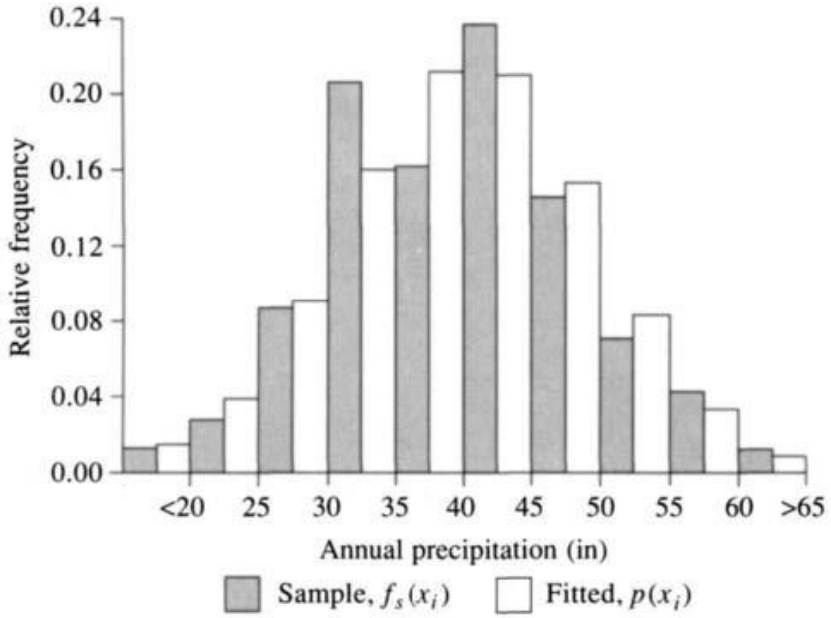
$$= F(35) - F(30)$$

$$= 0.301 - 0.144$$

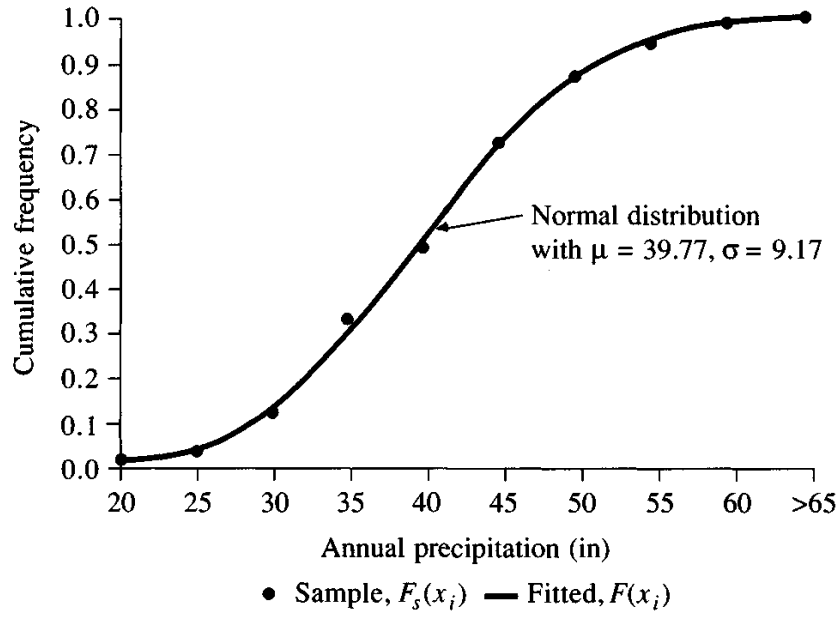
$$= 0.158$$

and similarly computed values for the other intervals are shown in column 7.

The relative frequency functions $f_s(x_i)$ and $p(x_i)$ from Table 11.4.1 are plotted in Fig. 11.4.3(a), and the cumulative frequency and probability distribution functions $F_s(x_i)$ and $F(x)$ in Fig. 11.4.3(b). From the similarity of the two functions



(*a*) Relative frequency function.



(*b*) Cumulative frequency function.

FIGURE 11.4.3
Frequency functions for a normal distribution fitted to annual precipitation in College Station, Texas (Example 11.4.3).

shown in each plot, it is apparent that the normal distribution fits these annual precipitation data very well.

To check the goodness of fit, the χ^2 test statistic is calculated by (11.4.4). For $i = 4$,

$$\frac{n[f_s(x_4) - p(x_4)]^2}{p(x_4)} = \frac{69 \times (0.20290 - 0.15777)^2}{0.15777}$$
$$= 0.891$$

as shown in column 8 of Table 11.4.1. The total of the values in column 8 is $\chi_c^2 = 2.377$. The value of $\chi^2_{\nu,\,1-\alpha}$ for a cumulative probability of $1-\alpha = 0.95$ and degrees of

freedom $\nu = m - p - 1 = 10 - 2 - 1 = 7$ is $\chi^2_{7,0.95} = 14.1$ (Abramowitz and Stegun, 1965). Since this value is greater than χ^2_c, the null hypothesis (the distribution fits the data) cannot be rejected at the 95 percent confidence level; the fit of the normal distribution to the College Station annual precipitation data is accepted. If the distribution had fitted poorly, the values of $f_s(x_i)$ and $p(x_i)$ would have been quite different from one another, resulting in a value of χ^2_c larger than 14.1, in which case the null hypothesis would have been rejected.

11.5 PROBABILITY DISTRIBUTIONS FOR HYDROLOGIC VARIABLES

In Sec. 11.4, the normal distribution was used to describe annual precipitation at College Station, Texas. Although this distribution fits this set of data particularly well, observations of other hydrologic variables follow different distributions. In this section, a selection of probability distributions commonly used for hydrologic variables is presented, and examples of the types of variables to which these distributions have been applied are given. Table 11.5.1 summarizes, for each distribution, the probability density function and the range of the variable, and gives equations for estimating the distribution's parameters from sample moments.

Normal Distribution

The normal distribution arises from the *central limit theorem*, which states that if a sequence of random variables X_i are independently and identically distributed with mean μ and variance σ^2, then the distribution of the sum of n such random variables, $Y = \sum_{i=1}^{n} X_i$, tends towards the normal distribution with mean $n\mu$ and variance $n\sigma^2$ as n becomes large. The important point is that this is true no matter what the probability distribution function of X is. So, for example, the probability distribution of the sample mean $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ can be approximated as normal with mean μ and variance $(1/n)^2 n\sigma^2 = \sigma^2/n$ no matter what the distribution of x is. Hydrologic variables, such as annual precipitation, calculated as the sum of the effects of many independent events tend to follow the normal distribution. The main limitations of the normal distribution for describing hydrologic variables are that it varies over a continuous range $[-\infty, \infty]$, while most hydrologic variables are nonnegative, and that it is symmetric about the mean, while hydrologic data tend to be skewed.

Lognormal Distribution

If the random variable $Y = \log X$ is normally distributed, then X is said to be lognormally distributed. Chow (1954) reasoned that this distribution is applicable to hydrologic variables formed as the products of other variables since if $X = X_1 X_2 X_3 \ldots X_n$, then $Y = \log X = \sum_{i=1}^{n} \log X_i = \sum_{i=1}^{n} Y_i$, which tends to the normal distribution for large n provided that the X_i are independent and identically distributed. The lognormal distribution has been found to describe the distribution of hydraulic conductivity in a porous medium (Freeze, 1975),

TABLE 11.5.1
Probability distributions for fitting hydrologic data

Distribution	Probability density function	Range	Equations for parameters in terms of the sample moments
Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$-\infty \leq x \leq \infty$	$\mu = \bar{x},\ \sigma = s_x$
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)$ where $y = \log x$	$x > 0$	$\mu_y = \bar{y},\ \sigma_y = s_y$
Exponential	$f(x) = \lambda e^{-\lambda x}$	$x \geq 0$	$\lambda = \frac{1}{\bar{x}}$
Gamma	$f(x) = \frac{\lambda^\beta x^{\beta-1} e^{-\lambda x}}{\Gamma(\beta)}$ where Γ = gamma function	$x \geq 0$	$\lambda = \frac{\bar{x}}{s_x^2}$ $\beta = \frac{\bar{x}^2}{s_x^2} = \frac{1}{CV^2}$

TABLE 11.5.1 (*cont.*)

Probability distributions for fitting hydrologic data

Distribution	Probability density function	Range	Equations for parameters in terms of the sample moments
Pearson Type III (three parameter gamma)	$f(x) = \dfrac{\lambda^{\beta}(x-\epsilon)^{\beta-1}e^{-\lambda(x-\epsilon)}}{\Gamma(\beta)}$	$x \geq \epsilon$	$\lambda = \dfrac{s_x}{\sqrt{\beta}}$, $\beta = \left(\dfrac{2}{C_s}\right)^2$ $\epsilon = \bar{x} - s_x\sqrt{\beta}$
Log Pearson Type III	$f(x) = \dfrac{\lambda^{\beta}(y-\epsilon)^{\beta-1}e^{-\lambda(y-\epsilon)}}{x\Gamma(\beta)}$ where $y = \log x$	$\log x \geq \epsilon$	$\lambda = \dfrac{s_y}{\sqrt{\beta}}$, $\beta = \left[\dfrac{2}{C_s(y)}\right]^2$ $\epsilon = \bar{y} - s_y\sqrt{\beta}$ (assuming $C_s(y)$ is positive)
Extreme Value Type I	$f(x) = \dfrac{1}{\alpha}\exp\left[-\dfrac{x-u}{\alpha} - \exp\left(-\dfrac{x-u}{\alpha}\right)\right]$	$-\infty < x < \infty$	$\alpha = \dfrac{\sqrt{6}s_x}{\pi}$ $u = \bar{x} - 0.5772\alpha$

the distribution of raindrop sizes in a storm, and other hydrologic variables. The lognormal distribution has the advantages over the normal distribution that it is bounded ($X > 0$) and that the log transformation tends to reduce the positive skewness commonly found in hydrologic data, because taking logarithms reduces large numbers proportionately more than it does small numbers. Some limitations of the lognormal distribution are that it has only two parameters and that it requires the logarithms of the data to be symmetric about their mean.

Exponential Distribution

Some sequences of hydrologic events, such as the occurrence of precipitation, may be considered *Poisson processes*, in which events occur instantaneously and independently on a time horizon, or along a line. The time between such events, or *interarrival time*, is described by the exponential distribution whose parameter λ is the mean rate of occurrence of the events. The exponential distribution is used to describe the interarrival times of random shocks to hydrologic systems, such as slugs of polluted runoff entering streams as rainfall washes the pollutants off the land surface. The advantage of the exponential distribution is that it is easy to estimate λ from observed data and the exponential distribution lends itself well to theoretical studies, such as a probability model for the linear reservoir ($\lambda = 1/k$, where k is the storage constant in the linear reservoir). Its disadvantage is that it requires the occurrence of each event to be completely independent of its neighbors, which may not be a valid assumption for the process under study—for example, the arrival of a front may generate many showers of rain—and this has led investigators to study various forms of *compound Poisson processes*, in which λ is considered a random variable instead of a constant (Kavvas and Delleur, 1981; Waymire and Gupta, 1981).

Gamma Distribution

The time taken for a number β of events to occur in a Poisson process is described by the gamma distribution, which is the distribution of a sum of β independent and identical exponentially distributed random variables. The gamma distribution has a smoothly varying form like the typical probability density function illustrated in Fig. 11.2.1 and is useful for describing skewed hydrologic variables without the need for log transformation. It has been applied to describe the distribution of depth of precipitation in storms, for example. The gamma distribution involves the *gamma function* $\Gamma(\beta)$, which is given by $\Gamma(\beta) = (\beta-1)! = (\beta-1)(\beta-2)\ldots3\cdot2\cdot1$ for positive integer β, and in general by

$$\Gamma(\beta) = \int_0^\infty u^{\beta-1}e^{-u}\,du \tag{11.5.1}$$

(Abramowitz and Stegun, 1965). The two-parameter gamma distribution (parameters β and λ) has a lower bound at zero, which is a disadvantage for application to hydrologic variables that have a lower bound larger than zero.

Pearson Type III Distribution

The Pearson Type III distribution, also called the *three-parameter gamma distribution*, introduces a third parameter, the lower bound ϵ, so that by the method of moments, three sample moments (the mean, the standard deviation, and the coefficient of skewness) can be transformed into the three parameters λ, β, and ϵ of the probability distribution. This is a very flexible distribution, assuming a number of different shapes as λ, β, and ϵ vary (Bobee and Robitaille, 1977).

The Pearson system of distributions includes seven types; they are all solutions for $f(x)$ in an equation of the form

$$\frac{d[f(x)]}{dx} = \frac{f(x)(x-d)}{C_0 + C_1 x + C_2 x^2} \tag{11.5.2}$$

where d is the *mode* of the distribution (the value of x for which $f(x)$ is a maximum) and C_0, C_1, and C_2 are coefficients to be determined. When $C_2 = 0$, the solution of (11.5.2) is a Pearson Type III distribution, having a probability density function of the form shown in Table 11.5.1. For $C_1 = C_2 = 0$, a normal distribution is the solution of (11.5.2). Thus, the normal distribution is a special case of the Pearson Type III distribution, describing a nonskewed variable. The Pearson Type III distribution was first applied in hydrology by Foster (1924) to describe the probability distribution of annual maximum flood peaks. When the data are very positively skewed, a log transformation is used to reduce the skewness.

Log–Pearson Type III Distribution

If log X follows a Pearson Type III distribution, then X is said to follow a log–Pearson Type III distribution. This distribution is the standard distribution for frequency analysis of annual maximum floods in the United States (Benson, 1968), and its use is described in detail in Chap. 12. As a special case, when log X is symmetric about its mean, the log–Pearson Type III distribution reduces to the lognormal distribution.

The location of the bound ϵ in the log–Pearson Type III distribution depends on the skewness of the data. If the data are positively skewed, then log $X \geq \epsilon$ and

TABLE 11.5.2
Shape and mode location of the log–Pearson Type III distribution as a function of its parameters

Shape parameter β	$\lambda < -\ln 10$	$-\ln 10 < \lambda < 0$	$\lambda > 0$
$0 < \beta < 1$	No mode J-shaped	Minimum mode U-shaped	No mode Reverse J-shaped
$\beta > 1$	Unimodal	No mode Reverse J-shaped	Unimodal

Source: Bobee, 1975.

ϵ is a lower bound, while if the data are negatively skewed, $\log X \leq \epsilon$ and ϵ is an upper bound. The log transformation reduces the skewness of the transformed data and may produce transformed data which are negatively skewed from original data which are positively skewed. In that case, the application of the log–Pearson Type III distribution would impose an artificial upper bound on the data. Depending on the values of the parameters, the log–Pearson Type III distribution can assume many different shapes, as shown in Table 11.5.2 (Bobee, 1975).

As described previously, the log–Pearson Type III distribution was developed as a method of fitting a curve to data. Its use is justified by the fact that it has been found to yield good results in many applications, particularly for flood peak data. The fit of the distribution to data can be checked using the χ^2 test, or by using probability plotting as described in Chap. 12.

Extreme Value Distribution

Extreme values are selected maximum or minimum values of sets of data. For example, the annual maximum discharge at a given location is the largest recorded discharge value during a year, and the annual maximum discharge values for each year of historical record make up a set of extreme values that can be analyzed statistically. Distributions of the extreme values selected from sets of samples of any probability distribution have been shown by Fisher and Tippett (1928) to converge to one of three forms of *extreme value distributions,* called Types I, II, and III, respectively, when the number of selected extreme values is large. The properties of the three limiting forms were further developed by Gumbel (1941) for the Extreme Value Type I (EVI) distribution, Frechet (1927) for the Extreme Value Type II (EVII), and Weibull (1939) for the Extreme Value Type III (EVIII).

The three limiting forms were shown by Jenkinson (1955) to be special cases of a single distribution called the *General Extreme Value* (GEV) distribution. The probability distribution function for the GEV is

$$F(x) = \exp\left[-\left(1 - k\frac{x-u}{\alpha}\right)^{1/k}\right] \tag{11.5.3}$$

where k, u, and α are parameters to be determined.

The three limiting cases are (1) for $k = 0$, the Extreme Value Type I distribution, for which the probability density function is given in Table 11.5.1, (2) for $k < 0$, the Extreme Value Type II distribution, for which (11.5.3) applies for $(u + \alpha/k) \leq x \leq \infty$, and (3) for $k > 0$, the Extreme Value Type III distribution, for which (11.5.3) applies for $-\infty \leq x \leq (u + \alpha/k)$. In all three cases, α is assumed to be positive.

For the EVI distribution x is unbounded (Table 11.5.1), while for EVII, x is bounded from below (by $u + \alpha/k$), and for the EVIII distribution, x is similarly bounded from above. The EVI and EVII distributions are also known as the *Gumbel* and *Frechet* distributions, respectively. If a variable x is described by the EVIII distribution, then $-x$ is said to have a *Weibull* distribution.

REFERENCES

Abramowitz, M., and I. A. Stegun, *Handbook of Mathematical Functions,* Dover, New York, p. 932, 1965.

Benson, M. A., Uniform flood-frequency estimating methods for federal agencies, *Water Resour. Res.,* vol. 4, no. 5, pp. 891–908, 1968.

Bobee, B., The log–Pearson Type III distribution and its application in hydrology, *Water Resour. Res.,* vol. 11, no. 5, pp. 681–689, 1975.

Bobee, B. B., and R. Robitaille, The use of the Pearson Type 3 and log Pearson Type 3 distributions revisited, *Water Resour. Res.,* vol. 13, no. 2, pp. 427–443, 1977.

Chow, V. T., The log-probability law and its engineering applications, *Proc. Am. Soc. Civ. Eng.,* vol. 80, pp. 1–25, 1954.

Fisher, R. A., On the mathematical foundations of theoretical statistics, *Trans. R. Soc. London A,* vol. 222, pp. 309–368, 1922.

Fisher, R. A., and L. H. C. Tippett, Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proc. Cambridge Phil. Soc.,* vol. 24, part II, pp. 180–191, 1928.

Foster, H. A., Theoretical frequency curves and their application to engineering problems, *Trans. Am. Soc. Civ. Eng.,* vol. 87, pp. 142–173, 1924.

Frechet, M., Sur la loi de probabilite de l'ecart maximum ("On the probability law of maximum values"), *Annales de la societe Polonaise de Mathematique;* vol. 6, pp. 93–116, Krakow, Poland, 1927.

Freeze, R. A., A stochastic-conceptual analysis of one-dimensional groundwater flow in nonuniform homogenous media, *Water Resour. Res.,* vol. 11, no. 5, pp. 725–741, 1975.

Gumbel, E. J., The return period of flood flows, *The Annals of Mathematical Statistics,* vol. 12, no. 2, pp. 163–190, June 1941.

Haan, C. T., *Statistical Methods in Hydrology,* Iowa State Univ. Press, Ames, Iowa, 1977.

Jenkinson, A. F., The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quart. Jour. Roy. Met. Soc.,* vol. 81, pp. 158–171, 1955.

Kavvas M. L., and J. W. Delleur, A stochastic cluster model of daily rainfall sequences, *Water Resour. Res.,* vol. 17, no. 4, pp. 1151–1160, 1981.

Pearson, K., On the systematic fitting of curves to observations and measurements, *Biometrika,* vol. 1, no. 3, pp. 265–303, 1902.

Waymire, E., and V. K. Gupta, The mathematical structure of rainfall representations I. A review of the stochastic rainfall models; *Water Resour. Res.,* vol. 17, no. 5, pp. 1261–1294, 1981.

Weibull, W., A statistical theory of the strength of materials, *Ingeniors Vetenskaps Akademien* (The Royal Swedish Institute for Engineering Research), proceedings no. 51, pp. 5–45, 1939.

PROBLEMS

11.1.1 The annual precipitation data for College Station, Texas, from 1911 to 1979 are given in Table 11.1.1. Estimate from the data the probability that the annual precipitation will be greater than 50 in in any year. Calculate the probability that annual precipitation will be greater than 50 in in two successive years (*a*) by assuming annual precipitation is an independent process; (*b*) directly from the data. Do the data suggest there is any tendency for years of precipitation > 50 in to follow one another in College Station?

11.1.2 Solve Prob. 11.1.1 for precipitation less than 30 in. Is there a tendency for years of precipitation less than 30 in to follow each other more than independence of events from year to year would suggest?

11.3.1 Calculate the mean, standard deviation, and coefficient of skewness for College Station annual precipitation from 1960 to 1969. The data are given in Table 11.1.1.

11.3.2 Calculate the mean, standard deviation, and coefficient of skewness for College Station annual precipitation for the six 10-year periods beginning in 1920, 1930, 1940, 1950, 1960, 1970 (e.g., 1920-1929). Compare the values of these statistics for the six samples. Calculate the mean and standard deviation of the six sample means and their coefficient of variation. Repeat this exercise for the six sample standard deviations and the six coefficients of skewness. As measured by the coefficient of variation of each sample statistic, which of these three sample statistics (mean, standard deviation, or coefficient of skewness) varies most from sample to sample?

11.4.1 Prove that the mean μ of the exponential distribution $f(x) = \lambda e^{-\lambda x}$ is given by $\mu = 1/\lambda$.

11.4.2 Show that the maximum likelihood estimates of the parameters of the normal distribution are given by

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

11.4.3 Calculate the value of the maximum likelihood estimates of the parameters of the normal distribution fitted to College Station annual precipitation from 1970 to 1979. Use the formulas given in Prob. 11.4.2 above and the data given in Table 11.1.1. Compare the result with the moment estimates given in Example 11.3.1.

11.4.4 Calculate the value of the log-likelihood function of College Station annual precipitation from 1970 to 1979 with $\mu = 40.17$ in and $\sigma = 10.63$ in. Holding μ constant, recompute and plot the value of the log-likelihood function by varying σ in increments of 0.1 from 9.5 to 11.5. Determine the value of σ that maximizes the log-likelihood function.

11.4.5 Solve Example 11.1.1 in the text using the probabilities for events A and B calculated from a normal distribution with $\mu = 39.77$ in and $\sigma = 9.17$ in (as fitted to the College Station precipitation data in Example 11.4.3). Compare the results you obtain with those in Example 11.1.1. Which method do you think is more reliable?

11.4.6 A reservoir system near College Station, Texas, is experiencing a drought and it is determined that if next year's annual precipitation in the reservoir watershed is less than 35 in, a reduction in the reservoir water supplied for irrigation will be required during the following year. If the annual precipitation is less than 35 in for each of the next two years, a reduction in municipal water supply will also be required. Using the normal distribution fitted to the precipitation data in Example 11.4.3, calculate the probability that these supply reductions will be necessary. Do you think these probabilities are sufficiently high to justify warning the irrigation and municipal water users of possible supply reductions?

11.5.1 The Pearson system of distributions obeys the equation $d[f(x)]/dx = [f(x)(x - d)]/(C_0 + C_1x + C_2x^2)$ where d is the mode of the distribution [the value of x where $f(x)$ is maximized] and C_0, C_1, and C_2 are coefficients. By setting $C_2 = 0$, show that the Pearson Type III distribution is obtained.

11.5.2 In Prob. 11.5.1, set $C_1 = C_2 = 0$ and show that the normal distribution is obtained.

11.5.3 The demand on a city's water treatment and distribution system is rising to near system capacity because of a long period of hot, dry weather. Rainfall will avert a situation where demand exceeds system capacity. If the average time between rainfalls in this city at this time of year is 5 days, calculate the chance that

there will be no rain (*a*) for the next 5 days, (*b*) 10 days, (*c*) 15 days. Use the exponential distribution.

11.5.4 Data for the annual maximum discharge of the Guadalupe River at Victoria, Texas, are presented in Table 12.1.1. The statistics for the logarithms to base 10 of these data are $\bar{y} = 4.2743$ and $s_y = 0.3981$. Fit the lognormal distribution to these data. Plot the relative frequency and incremental probability functions, and the cumulative frequency and probability distribution functions of the data as shown in Fig. 11.4.3 (use a log scale for the Guadalupe River discharges).

11.5.5 Data for inflow to the site of the proposed Justiceburg reservoir are given in Table 15.P.5. Calculate the mean, standard deviation, and coefficient of skewness of the annual total inflows and fit a probability distribution to the data.