Chapter 4

Parametric Probability Distributions

4.1. BACKGROUND

4.1.1. Parametric vs. Empirical Distributions

In Chapter 3, methods for exploring and displaying variations in data sets were presented. These methods had at their core the expression of how, empirically, a particular set of data are distributed through their range. This chapter presents an approach to the summarization of data distributions that involves imposition of particular mathematical forms, called *parametric distributions*, to represent variations in the underlying data. These mathematical forms are theoretical constructs, and so yield idealizations of real data.

It is worth taking a moment to understand why we would want to force real data to fit an abstract mold. The question is worth considering because parametric distributions *are* abstractions. They will represent real data only approximately, although in many cases the approximation can be very good indeed. There are three ways in which employing parametric probability distributions may be useful.

- *Compactness*. Particularly when dealing with large data sets, repeatedly manipulating the raw data can be cumbersome, or even severely limiting. A well-fitting parametric distribution reduces the number of quantities required for characterizing properties of the data from the full n order statistics ($x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$) to a small number of distribution parameters.
- *Smoothing and interpolation*. Real data are subject to sampling variations that lead to gaps or rough spots in their empirical distributions. For example, in Figures 3.1 and 3.11a there are no maximum temperature values between 10°F and 16°F, although certainly maximum temperatures in this range can and do occur during January at Ithaca. A parametric distribution imposed on these data would represent the possibility of these temperatures occurring, as well as allowing estimation of their probabilities of occurrence.
- *Extrapolation*. Estimating probabilities for events outside the range of a particular data set requires assumptions about as-yet-unobserved behavior. Again referring to Figure 3.11a, the empirical cumulative probability associated with the coldest temperature, 9°F, was estimated as 0.0213 using the Tukey plotting position. The probability of a maximum temperature this cold or colder could be estimated as 0.0213, but nothing further can be said quantitatively about the probability of January maximum temperatures colder than 5°F or 0°F without the imposition of a probability model such as that produced by a parametric distribution.

The distinction is being drawn between empirical and parametric data representations, but it should be emphasized that use of parametric probability distributions is not independent of empirical considerations. In particular, before embarking on the representation of data using parametric functions, we must

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00004-3

decide among candidate distribution forms, fit parameters of the chosen distribution, and check that the resulting function does, indeed, provide a reasonable fit. All three of these steps require use of real data.

4.1.2. What Is a Parametric Distribution?

A parametric distribution is an abstract mathematical form, or characteristic shape, or a family of characteristic shapes. Some of these mathematical forms arise naturally as a consequence of certain kinds of data-generating processes, and when applicable these are especially plausible candidates for concisely representing variations in a set of data. Even when there is not a strong natural justification behind the choice of a particular parametric distribution, it may be found empirically that the distribution represents a set of data very well.

The specific nature of a parametric distribution is determined by particular values for entities called *parameters* of that distribution. For example, the Gaussian (or "normal") distribution has as its characteristic shape the familiar symmetric bell. However, merely asserting that a particular batch of data, say average September temperatures at a location of interest, is well represented by the Gaussian distribution is not very informative about the nature of the data, without specifying *which* Gaussian distribution represents the data. There are infinitely many particular examples of the Gaussian distribution, corresponding to all possible values of the two distribution parameters μ and σ. But knowing, for example, that the monthly temperature for September is well represented by the Gaussian distribution with $\mu = 60°\text{F}$ and $\sigma = 2.5°\text{F}$ conveys a large amount of information about the nature and magnitudes of the variations of September temperatures at that location.

4.1.3. Parameters vs. Statistics

There is potential for confusion between distribution parameters and *sample statistics*. Distribution parameters are abstract characteristics of a particular parametric distribution. They succinctly represent underlying population, or data-generating process, properties. By contrast, a statistic is any quantity computed from a sample of data. Usually, the notation for sample statistics uses Roman (i.e., ordinary) letters, and parameters are typically written using Greek letters.

The confusion between parameters and statistics arises because, for some common parametric distributions, certain sample statistics are good estimators for the distribution parameters. For example, the sample standard deviation, s (Equation 3.6), a statistic, can be confused with the parameter σ of the Gaussian distribution because the two often are equated when finding a particular Gaussian distribution to best match a data sample. Distribution parameters are estimated (fitted) using sample statistics. However, it is not always the case that the fitting process is as simple as that for the Gaussian distribution, where usually the sample mean is equated to the parameter μ and the sample standard deviation is equated to the parameter σ.

4.1.4. Discrete vs. Continuous Distributions

There are two distinct types of parametric distributions, corresponding to different types of data, or random variables. *Discrete distributions* describe random quantities (i.e., the data of interest) that can take on only particular values. That is, the allowable values are finite, or at least countably infinite. For example, a *discrete random variable* might take on only the values 0 or 1; or any of the nonnegative integers; or one of the colors red, yellow, or blue. A *continuous random variable* typically can take on any value within a specified

range of the real numbers. For example, a continuous random variable might be defined on the real numbers between 0 and 1, or the nonnegative real numbers, or, for some distributions, the entire real line.

Strictly speaking, using a *continuous distribution* to represent observable data implies that the underlying observations are known to an arbitrarily large number of significant figures. Of course this is never true, but it is convenient and not too inaccurate to represent as continuous those variables that are continuous conceptually but reported discretely. Temperature and precipitation are two obvious examples that really range over some portion of the real number line, but which are usually reported to discrete multiples of 1°F and 0.01 in. in the United States. Little is lost when treating these discrete observations as samples from continuous distributions.

4.2. DISCRETE DISTRIBUTIONS

A large number of parametric distributions exist that are applicable to discrete random variables. Many of these are listed in the encyclopedic volume by Johnson et al. (1992), together with results concerning their properties. Only five of these, the binomial distribution, the geometric distribution, the negative binomial distribution, the multinomial distribution, and the Poisson distribution, are presented here.

4.2.1. Binomial Distribution

The *binomial distribution* is one of the simplest parametric distributions, and therefore is employed often in textbooks to illustrate the use and properties of parametric distributions more generally. This distribution pertains to outcomes of situations where, on some number of occasions (sometimes called "trials"), one or the other of two MECE (mutually exclusive and collectively exhaustive) events will occur. Classically the two events have been called "success" and "failure," but these are arbitrary labels. More generally, one of the events (say, the success) is assigned the number 1, and the other (the failure) is assigned the number zero.

The random variable of interest, X, is the number of event occurrences (given by the sum of 1s and 0s) in some number of trials. The number of trials, N, can be any positive integer, and the variable X can take on any of the nonnegative integer values from 0 (if the event of interest does not occur at all in the N trials) to N (if the event occurs on each occasion). The binomial distribution can be used to calculate probabilities for each of these $N + 1$ possible values of X if two conditions are met: (1) the probability of the event occurring does not change from trial to trial (i.e., the occurrence probability is *stationary*), and (2) the outcomes on each of the N trials are mutually independent. These conditions are rarely strictly met, but real situations can be close enough to this ideal that the binomial distribution provides sufficiently accurate representations.

One implication of the first restriction, relating to constant occurrence probability, is that events whose probabilities exhibit regular cycles must be treated carefully. For example, the event of interest might be thunderstorm or dangerous lightning occurrence, at a location where there is a diurnal or annual variation in the probability of the event. In cases like these, subperiods (e.g., hours or months, respectively) with approximately constant occurrence probabilities usually would be analyzed separately.

The second necessary condition for applicability of the binomial distribution, relating to event independence, is often more troublesome for atmospheric data. For example, the binomial distribution usually would not be directly applicable to daily precipitation occurrence or nonoccurrence. As illustrated by Example 2.2, such events often exhibit substantial day-to-day dependence. For situations like this the binomial distribution can be generalized to a theoretical stochastic process called a Markov

chain, discussed in Section 10.2. On the other hand, the year-to-year statistical dependence in atmospheric behavior is usually weak enough that occurrences or nonoccurrences of an event in consecutive annual periods can be considered to be effectively independent (12-month climate forecasts would be much easier if they were not!). The first four examples in this chapter take advantage of this fact.

The usual first illustration of the binomial distribution is in relation to coin flipping. If the coin is fair, the probability of either heads or tails is 0.5, and does not change from one coin-flipping occasion (or, equivalently, from one coin) to the next. If $N > 1$ coins are flipped simultaneously or in sequence, the outcome on one of the coins does not affect the other outcomes. The coin-flipping situation thus satisfies all the requirements for description by the binomial distribution: dichotomous, independent events with constant probability.

Consider a game where $N = 3$ fair coins are flipped simultaneously, and we are interested in the number, X, of heads that result. The possible values of X are 0, 1, 2, and 3. These four values are a MECE partition of the sample space for X, and their probabilities must therefore sum to 1. In this simple example, you may not need to think explicitly in terms of the binomial distribution to realize that the probabilities for these four events are 1/8, 3/8, 3/8, and 1/8, respectively.

In the general case, probabilities for each of the $N + 1$ values of X are given by the *probability distribution function*, or *probability mass function*, for the binomial distribution,

$$\Pr\{X = x\} = \binom{N}{x} p^x (1-p)^{N-x}, \quad x = 0, 1, \ldots, N. \tag{4.1}$$

Here, consistent with the usage in Equation 3.18, the uppercase X indicates the random variable whose precise value is unknown, or has yet to be observed. The lowercase x denotes a specific, particular value that the random variable can take on. The binomial distribution has two parameters, N and p. The parameter p is the probability of occurrence of the event of interest (the success) on any one of the N independent trials. For a given pair of the parameters N and p, Equation 4.1 is a function associating a probability with each of the discrete values $x = 0, 1, 2, \ldots, N$, such that $\sum_x \Pr\{X = x\} = 1$. That is, the probability distribution function distributes probability over all events in the sample space. Note that the binomial distribution is unusual in that both of its parameters are conventionally represented by Roman letters.

The right-hand side of Equation 4.1 consists of two parts: a combinatorial part and a probability part. The combinatorial part specifies the number of distinct ways of realizing x success outcomes from a collection of N trials. It is pronounced "N choose x" and is computed according to

$$\binom{N}{x} = \frac{N!}{x!(N-x)!}. \tag{4.2}$$

By convention, $0! = 1$. For example, when tossing $N = 3$ coins, there is only one way that $x = 3$ heads can be achieved: all three coins must come up heads. Using Equation 4.2, "three choose three" is given by $3!/(3!0!) = (1 \cdot 2 \cdot 3)/(1 \cdot 2 \cdot 3 \cdot 1) = 1$. There are three ways in which $x = 1$ can be achieved: either the first, the second, or the third coin can come up heads, with the remaining two coins coming up tails; using Equation 4.2 we obtain $3!/(1!2!) = (1 \cdot 2 \cdot 3)/(1 \cdot 1 \cdot 2) = 3$.

The probability part of Equation 4.1 follows from the multiplicative law of probability for independent events (Equation 2.12). The probability of a particular sequence of exactly x independent event occurrences and $N - x$ nonoccurrences is simply p multiplied by itself x times, and then multiplied by $1 - p$ (the probability of nonoccurrence) $N - x$ times. The number of these particular sequences of exactly x event occurrences and $N - x$ nonoccurrences is given by the combinatorial part, for each x, so that the product of the combinatorial and probability parts in Equation 4.1 yields the probability for x event occurrences, regardless of their locations in the sequence of N trials.

Example 4.1. Binomial Distribution and the Freezing of Cayuga Lake, I

Consider the data in Table 4.1, which lists years during which the surface of Cayuga Lake, in central New York State, was observed to have frozen. Cayuga Lake is rather deep and will freeze only after a long period of exceptionally cold and cloudy weather. In any given winter, the lake surface either freezes or it does not. Whether or not the lake freezes in a given winter is essentially independent of whether or not it froze in recent years. Unless there has been appreciable climate change in the region over the past 200 years, the probability that the lake will freeze in a given year has been effectively constant through the period of the data in Table 4.1. This assumption is increasingly questionable as the planet progressively warms, but if we can assume near-stationarity of the annual freezing probability, p, we expect the binomial distribution to provide a good statistical description of the freezing of this lake.

In order to use the binomial distribution as a representation of the statistical properties of the lake-freezing data, we need to *fit the distribution* to the data. Fitting the distribution simply means finding particular values for the distribution parameters, p and N in this case, for which Equation 4.1 will behave as much as possible like the data in Table 4.1. The binomial distribution is somewhat unique in that the parameter N depends on the question we want to ask, rather than on the data per se. If we want to compute the probability of the lake freezing next winter, or in any single winter in the future, $N = 1$. (The special case of Equation 4.1 with $N = 1$ is called the *Bernoulli distribution*, and one realization of a success or failure is called a *Bernoulli trial*.) If we want to compute probabilities for the lake freezing at least once during some decade in the future, $N = 10$.

The binomial parameter p in this application is the probability that the lake freezes in any given year. It is natural to estimate this probability using the relative frequency of the freezing events in the data. This is a straightforward task here, except for the small complication of not knowing exactly when the climate record starts. The written record clearly starts no later than 1796, but probably began some years before that. Suppose that the data in Table 4.1 represent a 230-year record. The 10 observed freezing events then lead to the relative frequency estimate for the binomial p of $10/230 = 0.0435$.[1]

We are now in a position to use Equation 4.1 to compute probabilities for a variety of events relating to the freezing of this lake. The simplest kinds of events to work with have to do with the lake freezing exactly a specified number of times, x, in a specified number of years, N. For example, the probability of the lake freezing exactly once in 10 years is

TABLE 4.1 Years in which Cayuga Lake Has Frozen, as of 2018

1796	1904
1816	1912
1856	1934
1875	1961
1884	1979

1. Cayuga lake nearly froze again in 2015. The 8 March visible satellite image from that year shows approximately 1/3 of the lake surface as open water. Incorporating this event into the data set of Table 4.1 would yield the estimate $p \approx 11/230 = 0.0478$ for the annual freezing probability.

$$\Pr\{X=1\} = \binom{10}{1}(.0435)^1(1-.0435)^{10-1} = \frac{10!}{1!9!}(.0435)(.9565)^9 = 0.292. \tag{4.3}$$

In reality the result in Equation 4.3 is likely an overestimate, because of the ongoing climate warming. ◇

Example 4.2. Binomial Distribution and the Freezing of Cayuga Lake, II

A somewhat harder class of events to deal with is exemplified by the problem of calculating the probability that the lake freezes at least once in 10 years. It is clear from Equation 4.3 that this probability will be no smaller than 0.292, since the probability for the compound event will be given by the sum of the probabilities $\Pr\{X=1\}+\Pr\{X=2\}+\cdots+\Pr\{X=10\}$. This result follows from Equation 2.5, and the fact that these events are mutually exclusive: the lake cannot freeze both exactly once and exactly twice in the same decade.

The brute-force approach to this problem is to calculate all 10 probabilities in the sum and then add them up. However, this approach is rather tedious, and quite a bit of effort can be saved by giving the problem a bit more thought. Consider that the sample space here is composed of 11 MECE events: that the lake freezes exactly 0, 1, 2, …, or 10 times in a decade. Since the probabilities for these 11 events must sum to 1, it is much easier to proceed using

$$\Pr\{X \geq 1\} = 1 - \Pr\{X=0\} = 1 - \frac{10!}{0!10!}(.0435)^0(.9565)^{10} = 0.359. \tag{4.4}$$

◇

It is worth noting that the binomial distribution can be applied to situations that are not intrinsically binary, through a suitable redefinition of events. For example, temperature is not intrinsically binary and is not even intrinsically discrete. However, for some applications it is of interest to consider the probability of frost, that is, $\Pr\{T \leq 32°\text{F}\}$. Together with the probability of the complementary event, $\Pr\{T > 32°\text{F}\}$, the situation is one concerning dichotomous events, and therefore could be a candidate for representation using the binomial distribution.

4.2.2. Geometric Distribution

The *geometric distribution* is related to the binomial distribution, describing a different aspect of the same data-generating situation. Both distributions pertain to a collection of independent trials in which one or the other of a pair of dichotomous events occurs. The trials are independent in the sense that the probability of the "success" occurring, p, does not depend on the outcomes of previous trials, and the sequence is stationary in the sense that p does not change over the course of the sequence (as a consequence of, e.g., an annual cycle or a changing climate). For the geometric distribution to be applicable, the collection of trials must occur in a sequence.

The binomial distribution pertains to probabilities that particular numbers of successes will be realized in a fixed number of trials. The geometric distribution specifies probabilities for the number of trials that will be required to observe the next success. For the geometric distribution, this number of trials is the random variable X, and the probabilities corresponding to its possible values are given by the probability distribution function of the geometric distribution,

$$\Pr\{X=x\} = p\,(1-p)^{x-1}\ ,\ x=1,2,\ldots \tag{4.5}$$

Here X can take on any positive integer value, since at least one trial will be required in order to observe a success, and it is possible (although vanishingly probable) that we would have to wait indefinitely for this

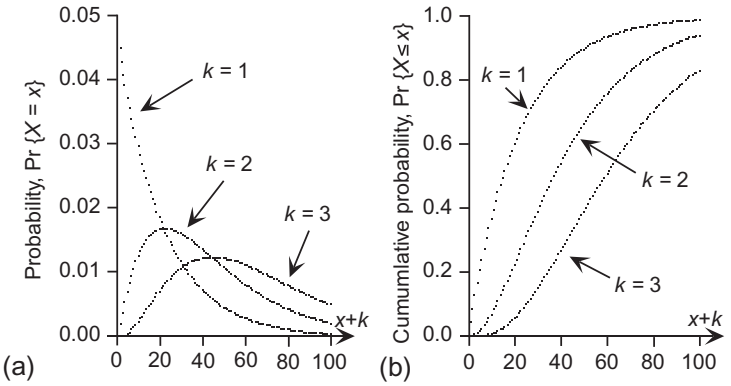

FIGURE 4.1 Probability distribution functions (a), and cumulative probability distribution functions (b), for the waiting time $x + k$ years for Cayuga Lake to freeze k times, using the negative binomial distribution, Equation 4.6.

outcome. Equation 4.5 can be viewed as an application of the multiplicative law of probability for independent events, as it multiplies the probability for a success by the probability of observing a sequence of $x-1$ consecutive failures. The function $k=1$ in Figure 4.1a shows an example geometric probability distribution, for the Cayuga Lake freezing probability $p=0.0435$.

Usually the geometric distribution is applied to trials that occur consecutively through time, so it is sometimes called the *waiting distribution*. The distribution has been used to describe lengths of weather regimes or spells. One application of the geometric distribution is description of sequences of dry time periods (where we are waiting for a wet event) and wet periods (during which we are waiting for a dry event), when the time dependence of events follows the first-order Markov process (Waymire and Gupta 1981; Wilks 1999a), described in Section 10.2.

4.2.3. Negative Binomial Distribution

The *negative binomial distribution* is closely related to the geometric distribution, although this relationship is not indicated by its name, which comes from a technical derivation with parallels to a similar derivation for the binomial distribution. The probability distribution function for the negative binomial distribution is defined for nonnegative integer values of the random variable x,

$$\Pr\{X=x\}=\frac{\Gamma(k+x)}{x!\Gamma(k)}p^k(1-p)^x, \quad x=0,1,2,\ldots \tag{4.6}$$

The distribution has two real-valued parameters, p, $0<p<1$ and k, $k>0$. For integer values of k the negative binomial distribution is called the *Pascal distribution* and has an interesting interpretation as an extension of the geometric distribution of waiting times for the first success in a sequence of independent Bernoulli trials with probability p. In this case, the negative binomial X pertains to the number of failures until the kth success, so that $x + k$ is the total waiting time required to observe the kth success.

The notation $\Gamma(k)$ on the right-hand side of Equation 4.6 indicates a standard mathematical function known as the *gamma function*, defined by the definite integral

TABLE 4.2 Values of the Gamma Function, $\Gamma(k)$ (Equation 4.7), for $1.00 \le k \le 1.99$

k	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	1.0000	0.9943	0.9888	0.9835	0.9784	0.9735	0.9687	0.9642	0.9597	0.9555
1.1	0.9514	0.9474	0.9436	0.9399	0.9364	0.9330	0.9298	0.9267	0.9237	0.9209
1.2	0.9182	0.9156	0.9131	0.9108	0.9085	0.9064	0.9044	0.9025	0.9007	0.8990
1.3	0.8975	0.8960	0.8946	0.8934	0.8922	0.8912	0.8902	0.8893	0.8885	0.8879
1.4	0.8873	0.8868	0.8864	0.8860	0.8858	0.8857	0.8856	0.8856	0.8857	0.8859
1.5	0.8862	0.8866	0.8870	0.8876	0.8882	0.8889	0.8896	0.8905	0.8914	0.8924
1.6	0.8935	0.8947	0.8959	0.8972	0.8986	0.9001	0.9017	0.9033	0.9050	0.9068
1.7	0.9086	0.9106	0.9126	0.9147	0.9168	0.9191	0.9214	0.9238	0.9262	0.9288
1.8	0.9314	0.9341	0.9368	0.9397	0.9426	0.9456	0.9487	0.9518	0.9551	0.9584
1.9	0.9618	0.9652	0.9688	0.9724	0.9761	0.9799	0.9837	0.9877	0.9917	0.9958

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt. \tag{4.7}$$

In general, the gamma function must be evaluated numerically (e.g., Abramowitz and Stegun, 1984; Press et al., 1986) or approximated using tabulated values, such as those given in Table 4.2. It satisfies the factorial recurrence relationship,

$$\Gamma(k+1) = k\,\Gamma(k), \tag{4.8}$$

allowing Table 4.2 to be extended indefinitely. For example, $\Gamma(3.50) = (2.50)\ \Gamma(2.50) = (2.50)(1.50)\ \Gamma(1.50) = (2.50)(1.50)(0.8862) = 3.323$. Similarly, $\Gamma(4.50) = (3.50)\ \Gamma(3.50) = (3.50)\ (3.323) = 11.631$. The gamma function is also known as the *factorial function*, the reason for which is especially clear when its argument is an integer (e.g., in Equation 4.6 when k is an integer), that is, $\Gamma(k+1) = k!$

With this understanding of the gamma function, it is straightforward to see the connection between the negative binomial distribution with integer k as a waiting distribution for k successes, and the geometric distribution (Equation 4.5) as a waiting distribution for the first success, in a sequence of independent Bernoulli trials with success probability p. Since X in Equation 4.6 is the number of failures before observing the kth success, and the total number of trials to achieve k successes will be $x + k$, then for $k = 1$, Equations 4.5 and 4.6 pertain to the same situation. The numerator in the first factor on the right-hand side of Equation 4.6 is $\Gamma(x+1) = x!$, canceling the $x!$ in the denominator. Realizing that $\Gamma(1) = 1$ (see Table 4.2), Equation 4.6 reduces to Equation 4.5 except that Equation 4.6 pertains to $k = 1$ additional trial since it also includes that $k = 1$st success.

Example 4.3. Negative Binomial Distribution and the Freezing of Cayuga Lake, III

Assuming again that the freezing of Cayuga Lake is well represented statistically by a series of annual Bernoulli trials with $p = 0.0435$, what can be said about the probability distributions for the number of years required to observe k winters in which the lake freezes? As noted earlier, these probabilities will be those pertaining to X in Equation 4.6.

Figure 4.1a shows three of these negative binomial distributions, for $k = 1$, 2, and 3, shifted to the right by k years in order to show the distributions of waiting times, $x + k$. That is, the leftmost points in the three functions in Figure 4.1a all correspond to $X = 0$ in Equation 4.6. For $k = 1$ the probability distribution function is the same as for the geometric distribution (Equation 4.5), and the figure shows that the probability of freezing in the next year is simply the Bernoulli $p = 0.0435$. The probabilities that year $x + 1$ will be the next freezing event decrease smoothly at a fast enough rate that probabilities for the first freeze being more than a century away are quite small. It is impossible for the lake to freeze $k = 2$ times before next year, so the first probability plotted in Figure 4.1a for $k = 2$ is at $x + k = 2$ years, and this probability is $p^2 = 0.0435^2 = 0.0019$. These probabilities rise through the most likely waiting time for two freezes at $x + 2 = 23$ years before falling again, although there is a nonnegligible probability that the lake still will not have frozen twice within a century. When waiting for $k = 3$ freezes, the probability distribution of waiting times is flattened more and shifted even further into the future.

An alternative way of viewing these distributions of waiting times is through their cumulative probability distribution functions,

$$\Pr\{X \le x\} = \sum_{t=0}^{x} \Pr\{X = t\}, \tag{4.9}$$

which are plotted in Figure 4.1b. Here all the probabilities for waiting times t less than or equal to a waiting time x of interest have been summed, analogously to Equation 3.18 for the empirical *cumulative distribution function* (CDF). For $k = 1$, the CDF rises rapidly at first, indicating that the probability of the first freeze occurring within the next few decades is quite high, and that it is nearly certain that the lake will freeze next within a century (assuming that the annual freezing probability p is stationary so that, e.g., it is not decreasing through time as a consequence of a changing climate). These functions rise more slowly for the waiting times for $k = 2$ and $k = 3$ freezes; and indicate a probability around 0.93 that the lake will freeze at least twice, and a probability near 0.82 that the lake will freeze at least three times, during the next century, again assuming that the climate is stationary. ◇

Use of the negative binomial distribution is not limited to integer values of the parameter k, and when k is allowed to take on any positive value the distribution may be appropriate for flexibly describing variations in data on counts. For example, the negative binomial distribution has been used (in slightly modified form) to represent the distributions of spells of consecutive wet and dry days (Wilks 1999a), and annual numbers of landfalling Atlantic hurricanes (Hall and Jewson 2008) in a way that is more flexible than Equation 4.5 because values of k different from 1 produce different shapes for the distribution, as in Figure 4.1a. In general, appropriate parameter values must be determined by the data to which the distribution will be fit. That is, specific values for the parameters p and k must be determined that will allow Equation 4.6 to look as much as possible like the empirical distribution of the data that it will be used to represent.

The simplest way to find appropriate values for the parameters in the more general situation of noninteger k, that is, to fit the distribution, is to use the *method of moments*. To use the method of moments we mathematically equate the sample moments and the distribution (or population) moments. Since there are two parameters, it is necessary to use two distribution moments to define them. The first moment is the mean and the second moment is the variance. In terms of the distribution parameters, the mean of the negative binomial distribution is $\mu = k(1-p)/p$, and the variance is $\sigma^2 = k(1-p)/p^2$. Estimating p and k using the method of moments involves simply setting these expressions equal to the corresponding sample moments and solving the two equations simultaneously for the parameters.

That is, each data value x is an integer, and the mean and variance of these x's are calculated, and substituted into the equations

$$p = \frac{\overline{x}}{s^2}, \tag{4.10a}$$

and

$$k = \frac{\overline{x}^2}{s^2 - \overline{x}}. \tag{4.10b}$$

4.2.4. Multinomial Distribution

The *multinomial distribution* extends the binomial distribution to situations where the MECE partition of the sample space consists of more than two discrete events. In common with the binomial distribution, the multinomial distribution relates to a sequence or collection of N independent trials, but on each of these trials one and only one of $K > 2$ outcomes X_k, $k = 1, \ldots, K$, occurs. The probabilities p_k, $k = 1, \ldots, K$ for each of these outcomes are unchanging (i.e., the process is stationary), and of course $\sum_k p_k = 1$.

The probability distribution function for the multinomial distribution is

$$\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_K = x_k\} = \frac{N!}{x_1!x_2!\ldots x_K!} p_1^{x_1} p_2^{x_2} \ldots p_K^{x_K}, \tag{4.11}$$

which distributes probability over the K outcome variables. Note that for $K = 2$, Equation 4.11 reduces to the probability distribution function for the binomial distribution (Equation 4.1), and that each of the X_k individually follow the binomial distribution with parameters p_k and N.

Because the K outcomes x_k must sum to N, they can jointly be represented geometrically on a $K-1$-dimensional hyperplane. Binomial outcomes, for example, can be represented on the real line (i.e., in one dimension), since either one of the numbers of the complementary events "success" and "failure" is sufficient to define the other. For $K = 3$ the outcomes can be jointly represented on a two-dimensional surface (see Section 9.4.9 for an example). Distribution fitting (i.e., estimation of the p_ks) is most easily achieved by substituting the respective outcome relative frequencies.

4.2.5. Poisson Distribution

The *Poisson distribution* describes the numbers of discrete events occurring in a series, or a sequence, and so pertains to data on counts that can take on only nonnegative integer values. Usually the sequence is understood to be in time, for example, the occurrence of storms in a particular geographic region over the course of a year. However, it is also possible to apply the Poisson distribution to counts of events occurring in one or more spatial dimensions, such as the number of gasoline stations along a particular stretch of highway, or the distribution of hailstones over a small area.

Poisson events occur randomly, but at a constant average rate. That is, the average rate at which Poisson events are generated is stationary. The individual events being counted must be independent, in the sense that their occurrences do not depend on whether or how many other events may have occurred elsewhere in nonoverlapping portions of the sequence. Given the average rate of event occurrence, the probabilities of particular numbers of events in a given interval depend only on the size of the

interval over which events will be counted. A sequence of such events is sometimes said to have been generated by a *Poisson process*. As was the case for the binomial distribution, strict adherence to this independence condition is often difficult to demonstrate in atmospheric data, but the Poisson distribution can still yield a useful representation if the degree of dependence is not too strong. Ideally, Poisson events should be rare enough that the probability of more than one occurring simultaneously is very small. One way of motivating the Poisson distribution mathematically is as the limiting case of the binomial distribution, as p approaches zero and N approaches infinity.

The Poisson distribution has a single parameter, μ, that specifies the average occurrence rate. The Poisson parameter is sometimes called the *intensity* and has physical dimensions of occurrences per unit time. The probability distribution function for the Poisson distribution is

$$\Pr\{X = x\} = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, 2, \ldots, \tag{4.12}$$

which associates probabilities with all possible numbers of occurrences, X, from zero to infinitely many. Here $e \approx 2.718$ is the base of the natural logarithms. The sample space for Poisson events therefore contains (countably) infinitely many elements. Clearly the summation of Equation 4.12 for x running from zero to infinity must be convergent and equal to 1. The probabilities associated with very large numbers of counts are vanishingly small, since the denominator in Equation 4.12 is $x!$

To use the Poisson distribution it must be fit to a sample of data. Again, fitting the distribution means finding the specific value for the single parameter μ that makes Equation 4.12 behave as similarly as possible to a data set at hand. For the Poisson distribution, a good way to estimate the parameter μ is by using the method of moments. Fitting the Poisson distribution is thus especially easy, since its one parameter is the mean number of occurrences per unit time, which can be estimated directly as the sample average of the number occurrences per unit time.

Example 4.4. Poisson Distribution and Annual U.S. Hurricane Landfalls

The Poisson distribution is a natural and commonly used statistical model for representing hurricane occurrence statistics (e.g., Parisi and Lund 2008). Consider the Poisson distribution in relation to the annual number of hurricanes making landfall on the U.S. coastline, from Texas through Maine, for 1899–1998, shown as the dashed histogram in Figure 4.2. During the 100 years covered by these data, 170 hurricanes made landfall on the U.S. coastline (Neumann et al. 1999). The counts range from zero U.S. hurricane landfalls in 16 of the 100 years, through six U.S. hurricane landfalls in two of the years (1916 and 1985). The average, or mean, rate of U.S. hurricane landfall occurrence is simply 170/100 = 1.7 landfalls/year, so this average is the method-of-moments estimate of the Poisson intensity for these data. Having fit the distribution by estimating a value for its parameter, the Poisson distribution can be used to compute probabilities that particular numbers of hurricanes will make landfall on the U.S. coastline annually. The first 8 of these probabilities (pertaining to zero through seven hurricane landfalls per year) are plotted in the form of the solid histogram in Figure 4.2.

The Poisson distribution allocates probability smoothly (within the limitation that the data are discrete) among the possible outcomes, with the most probable numbers of landfalls being near the mean rate of 1.7 per year. The distribution of the data shown by the dashed histogram resembles that of the fitted Poisson distribution but is more irregular, especially for the more active years, due at least in part to sampling variations. For example, there does not seem to be a physically based reason why five hurricanes per year should be less likely than six. Fitting a distribution to these data provides a sensible way to smooth out such variations, which is desirable if the irregular variations in the data histogram are not

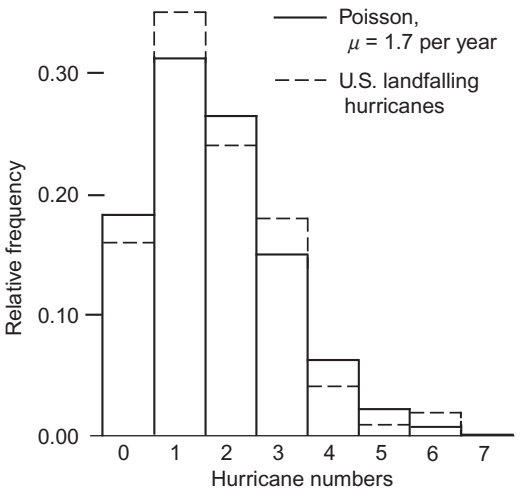

FIGURE 4.2 Histogram of numbers of U.S. landfalling hurricanes for 1899–1998 *(dashed)* and fitted Poisson distribution with $\mu = 1.7$ hurricanes/year *(solid)*.

physically meaningful. Similarly, using the Poisson distribution to summarize the data allows quantitative estimation of probabilities for large numbers of landfalls in a year. Even though none of the years in this 100-year record had more than six U.S. hurricane landfalls, even more active years are not physically impossible and the fitted Poisson distribution allows probabilities for such events to be estimated. For example, according to this Poisson model, the probability of seven U.S. hurricane landfalls occurring in a given year would be estimated as $\Pr\{X = 7\} = 1.7^7 e^{-1.7}/7! = 0.00149$. ◇

4.3. STATISTICAL EXPECTATIONS

4.3.1. Expected Value of a Random Variable

The *expected value* of a random variable or function of a random variable is simply the probability-weighted average of that variable or function. This weighted average is called the expected value, although we do not necessarily expect this outcome to occur in the informal sense of an "expected" event being likely. It can even happen that the statistical expected value is an impossible outcome. Statistical expectations are closely tied to probability distributions, since the distributions will provide the weights or weighting function for the weighted average. The ability to work easily with statistical expectations can be a strong motivation for choosing to represent data using parametric distributions rather than empirical distribution functions.

It is easiest to see expectations as probability-weighted averages in the context of a discrete probability distribution, such as the binomial. Conventionally, the expectation operator is denoted $E[]$, so that the expected value for a discrete random variable is

$$E[X] = \sum_x x \Pr\{X = x\}. \tag{4.13}$$

The equivalent notation $\langle X \rangle = E[X]$ is sometimes used for the expectation operator. The summation in Equation 4.13 is taken over all allowable values of X. For example, the expected value of X when X follows the binomial distribution is

TABLE 4.3 Expected Values (Means) and Variances for Probability Distribution Functions Described in Section 4.2, in Terms of Their Distribution Parameters

Distribution	Probability Distribution Function	$\mu = E[X]$	$\sigma^2 = Var[X]$
Binomial	Equation 4.1	Np	$Np(1-p)$
Geometric	Equation 4.5	$1/p$	$(1-p)/p^2$
Negative Binomial	Equation 4.6	$k(1-p)/p$	$k(1-p)/p^2$
Poisson	Equation 4.12	μ	μ

$$E[X] = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x}. \tag{4.14}$$

Here the allowable values of X are the nonnegative integers up to and including N, and each term in the summation consists of the specific value of the variable, x, multiplied by the probability of its occurrence from Equation 4.1.

The expectation $E[X]$ has a special significance, since it is the mean of the distribution of X. Distribution (or population, or data-generating process) means are conventionally denoted using the symbol μ. It is possible to analytically simplify Equation 4.14 to obtain, for the binomial distribution, the result $E[X] = Np$. Thus the mean of any binomial distribution is given by the product $\mu = Np$. Expected values for discrete probability distributions described in Section 4.2 are listed in Table 4.3, in terms of their distribution parameters. The U.S. hurricane landfall data in Figure 4.2 provide an example of the expected value $E[X] = 1.7$ landfalls being impossible to realize in any year.

4.3.2. Expected Value of a Function of a Random Variable

It can be very useful to compute expectations, or probability-weighted averages, of functions of random variables,
$E[g(x)]$. Since the expectation is a linear operator, expectations of functions of random variables have the following properties:

$$E[c] = c, \tag{4.15a}$$

$$E[c\, g_1(x)] = c\, E[g_1(x)], \tag{4.15b}$$

$$E\left[\sum_{j=1}^{J} g_j(x)\right] = \sum_{j=1}^{J} E\left[g_j(x)\right], \tag{4.15c}$$

where c is any constant and $g_j(x)$ is any function of x. Because the constant c does not depend on x, $E[c] = \sum_x c \Pr\{X = x\} = c \sum_x \Pr\{X = x\} = c \cdot 1 = c$. Equations 4.15a and 4.15b reflect the fact that constants can be factored out of the summations when computing expectations. Equation 4.15c expresses the important property that the expectation of a sum is equal to the sum of the separate expected values.

Use of the properties expressed in Equation 4.15 can be illustrated with the expectation of the function $g(x) = (x - \mu)^2$. The expected value of this function is called the *variance* and is conventionally denoted by σ^2. Applying the properties in Equations 4.15 to this expectation yields

$$\begin{aligned} Var[X] = E\left[(X-\mu)^2\right] &= \sum_x (x-\mu)^2 \Pr\{X=x\} \\ &= \sum_x \left(x^2 - 2\mu x + \mu^2\right) \Pr\{X=x\} \\ &= \sum_x x^2 \Pr\{X=x\} - 2\mu \sum_x x \Pr\{X=x\} + \mu^2 \sum_x \Pr\{X=x\} \\ &= E\left[X^2\right] - 2\mu E[X] + \mu^2 \cdot 1 \\ &= E\left[X^2\right] - \mu^2. \end{aligned} \tag{4.16}$$

Notice the similarity of the last equality on the first line of Equation 4.16 to the sample variance, given by the square of Equation 3.6. Similarly, the final equality in Equation 4.16 is analogous to the computational form for the sample variance, given by the square of Equation 3.31. Notice also that combining the first line of Equation 4.16 with the properties in Equation 4.15 yields

$$Var[c\, g(x)] = c^2\, Var[g(x)]. \tag{4.17}$$

Variances for four of the univariate discrete distributions described in Section 4.2 are listed in Table 4.3.

Example 4.5. Expected Value of a Function of a Binomial Random Variable

Table 4.4 presents the computation of statistical expectations for the binomial distribution with $N=3$ and $p=0.5$. These parameters correspond to the situation of simultaneously flipping three coins and counting $X=$the number of heads. The first column shows the possible outcomes of X, and the second column shows the probabilities for each of the outcomes, computed according to Equation 4.1.

The third column in Table 4.4 shows the individual terms in the probability-weighted average $E[X]=\sum_x [x \Pr(X=x)]$. Adding these four values yields $E[X]=1.5$, as would be obtained by multiplying the two distribution parameters $\mu = Np$, in Table 4.3

The fourth column in Table 4.4 similarly shows the construction of the expectation $E[X^2]=3.0$. We might imagine this expectation in the context of a hypothetical game, in which the player receives $\$X^2$, that is, nothing if zero heads come up, \$1 if one head comes up, \$4 if two heads come up, and \$9 if three heads come up. Over the course of many rounds of this game, the long-term average payout would be $E[X^2]=\$3.00$. An individual willing to pay more than \$3 to play this game would be either foolish or inclined toward taking risks.

TABLE 4.4 Binomial Probabilities for $N=3$ and $p=0.5$, and the Construction of the Expectations $E[X]$ and $E[X^2]$ as Probability-Weighted Averages

X	$\Pr(X=x)$	$x \cdot \Pr(X=x)$	$x^2 \cdot \Pr(X=x)$
0	0.125	0.000	0.000
1	0.375	0.375	0.375
2	0.375	0.750	1.500
3	0.125	0.375	1.125
		$E[X]=1.500$	$E[X^2]=3.000$

Notice that the final equality in Equation 4.16 can be verified for this particular binomial distribution using Table 4.4. Here $E[X^2]-\mu^2 = 3.0-(1.5)^2 = 0.75$, agreeing with $Var[X] = Np(1-p) = 3(0.5)(1-0.5) = 0.75$. ◇

4.4. CONTINUOUS DISTRIBUTIONS

Most atmospheric variables can take on any of a continuum of values. Temperature, precipitation amount, geopotential height, wind speed, and other quantities are at least conceptually not restricted to integer values of the physical units in which they are measured. Even though the nature of measurement and reporting systems is such that measurements are rounded to discrete values, the set of reportable values is large enough that most such variables can still be treated as continuous quantities.

Many continuous parametric distributions exist. Those used most frequently in the atmospheric sciences are discussed in subsequent sections. Encyclopedic information on these and many other continuous distributions can be found in Johnson et al. (1994, 1995).

4.4.1. Distribution Functions and Expected Values

The mathematics of probability for continuous variables are somewhat different, although analogous, to those for discrete random variables. In contrast to probability calculations for discrete distributions, which involve summation over a discontinuous probability distribution function (e.g., Equation 4.1), probability calculations for continuous random variables involve integration over continuous functions called *probability density functions* (PDFs). A PDF is sometimes referred to more simply as a *density*.

Conventionally, the PDF for a random variable X is denoted $f(x)$. Just as summation of a discrete probability distribution function over all possible values of the random quantity must equal 1, the integral of any PDF over all allowable values of x must equal 1:

$$\int_x f(x)dx = 1. \tag{4.18}$$

A function cannot be a PDF unless it satisfies this condition. Furthermore, a PDF $f(x)$ must be nonnegative for all values of x. No specific limits of integration have been included in Equation 4.18, because different probability densities are defined over different ranges of the random variable (i.e., have different *support*).

Probability density functions are the continuous parametric analogs of the familiar histogram (see Section 3.3.5) and of the nonparametric kernel density estimate (see Section 3.3.6). However, the meaning of the PDF is often initially confusing precisely because of the analogy with the histogram. In particular, the height of the density function $f(x)$, obtained when it is evaluated at a particular value of the random variable, is not in itself meaningful in the sense of defining a probability. The confusion arises because often it is not realized that probability is proportional to area, and not to height, in both the PDF and the histogram.

Figure 4.3 shows a hypothetical PDF, defined for a nonnegative random variable X. A PDF can be evaluated for a specific value of the random variable, say $x = 1$, but by itself $f(1)$ is not meaningful in terms of probabilities for X. In fact, since X varies continuously over some portion of

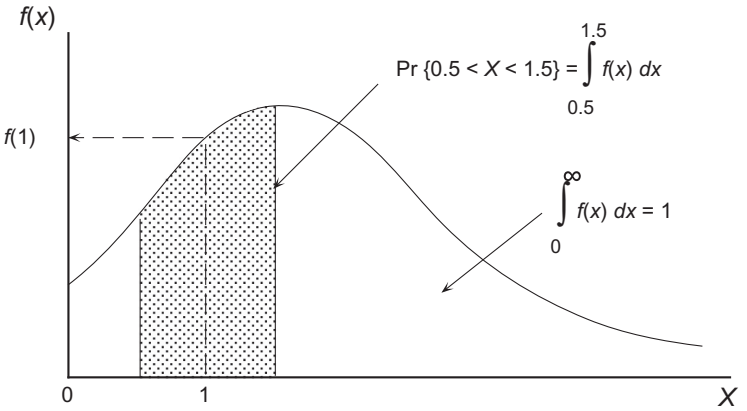

FIGURE 4.3 Hypothetical PDF $f(x)$ for a nonnegative random variable, X. Evaluation of $f(x)$ is not, by itself, meaningful in terms of probabilities for specific values of X. Probabilities are obtained by integrating portions of $f(x)$.

the real numbers, the probability of *exactly* $X = 1$ is infinitesimally small. It is meaningful, however, to think about and compute probabilities for values of a random variable in finite neighborhoods around $X = 1$. Figure 4.3 shows the probability of X being between 0.5 and 1.5 as the integral of the PDF between these limits.

An idea related to the PDF is that of the CDF. The CDF is a function of the random variable X, given by the integral of the PDF up to a particular value of x. Thus the CDF specifies probabilities that the random quantity X will not exceed particular values. It is therefore the continuous counterpart to the empirical CDF, Equation 3.18, and the discrete CDF, for example, Equation 4.9. Conventionally, CDFs are denoted $F(x)$:

$$F(x) = \Pr\{X \le x\} = \int_{X \le x} f(x)dx. \tag{4.19}$$

Again, specific integration limits have been omitted from Equation 4.19 to indicate that the integration is performed from the minimum allowable value of X to the particular value, x, that is the argument of the function. Since the values of $F(x)$ are probabilities, $0 \le F(x) \le 1$.

Equation 4.19 transforms a particular value of the random variable to a cumulative probability. The value of the random variable corresponding to a particular cumulative probability is given by the inverse of the CDF,

$$F^{-1}(p) = x(F), \tag{4.20}$$

where p is the cumulative probability. That is, Equation 4.20 specifies the upper limit of the integration in Equation 4.19 that will yield a particular cumulative probability $p = F(x)$. Since this inverse of the CDF specifies the data quantile corresponding to a particular probability, Equation 4.20 is also called the *quantile function*.

Statistical expectations are defined for continuous as well as for discrete random variables. As is the case for discrete variables, the expected value of a variable or a function is the probability-weighted average of that variable or function. Since probabilities for continuous random variables are computed by integrating their density functions, the expected value of a function of a random variable is given by the integral

TABLE 4.5 Expected Values (Means) and Variances for Continuous Probability Density Functions Described in This Section, in Terms of Their Parameters

Distribution	PDF	$E[X]$	$Var[X]$
Gaussian	Equation 4.24	μ	σ^2
Lognormal[1]	Equation 4.36	$\exp[\mu+\sigma^2/2]$	$(\exp[\sigma^2]-1)\exp[2\mu+\sigma^2]$
Zero-truncated Gaussian	Equation 4.39	Equation 4.41a	Equation 4.41b
Logistic	Equation 4.42	μ	$\sigma^2\pi^2/3$
Gamma	Equation 4.45	$\alpha\beta$	$\alpha\beta^2$
Exponential	Equation 4.52	β	β^2
Chi-square	Equation 4.54	ν	2ν
Pearson III	Equation 4.55	$\zeta+\alpha\beta$	$\alpha\beta^2$
Beta	Equation 4.58	$\alpha/(\alpha+\beta)$	$(\alpha\beta)/[(\alpha+\beta)^2(\alpha+\beta+1)]$
Gumbel[2]	Equation 4.57	$\zeta+\gamma\beta$	$\beta\pi/\sqrt{6}$
GEV[3]	Equation 4.63	$\zeta-\beta[1-\Gamma(1-\kappa)]/\kappa$	$\beta^2(\Gamma[1-2\kappa]-\Gamma^2[1-\kappa])/\kappa^2$
Weibull	Equation 4.66	$\beta\Gamma[1+1/\alpha]$	$\beta^2(\Gamma[1+2/\alpha]-\Gamma^2[1+1/\alpha])$
Mixed Exponential	Equation 4.78	$w\beta_1+(1-w)\beta_2$	$w\beta_1^2+(1-w)\beta_2^2+w(1-w)(\beta_1-\beta_2)^2$

[1]*For the lognormal distribution, μ and σ^2 refer to the mean and variance of the log-transformed variable $y=\ln(x)$.*
[2]*$\gamma=0.57721\ldots$ is Euler's constant.*
[3]*For the GEV the mean exists (is finite) only for $\kappa<1$, and the variance exists only for $\kappa<1/2$.*

$$E[g(x)]=\int_x g(x)f(x)\,dx. \tag{4.21}$$

Expectations of continuous random variables also exhibit the properties in Equations 4.15 and 4.17. For $g(x)=x$, $E[X]=\mu$ is the mean of the distribution whose PDF is $f(x)$. Similarly, the variance of a continuous variable is given by the expectation of the function $g(x)=(x-E[X])^2$,

$$\begin{aligned} Var[X]=E\left[(x-E[X])^2\right] &= \int_x (x-E[X])^2 f(x)\,dx \\ &= \int_x x^2 f(x)\,dx-(E[X])^2=E\left[X^2\right]-\mu^2. \end{aligned} \tag{4.22}$$

Note that, depending on the particular functional form of $f(x)$, some or all of the integrals in Equations 4.19, 4.21, and 4.22 may not be analytically computable, and for some distributions the integrals may not even exist.

Table 4.5 lists means and variances for the distributions to be described in this section, in terms of the distribution parameters.

4.4.2. Gaussian Distributions

The *Gaussian distribution* plays a central role in classical statistics and has many applications in the atmospheric sciences as well. It is sometimes also called the *normal* distribution, although this name

carries the unwarranted connotation that it is in some way universal, or that deviations from it are in some way unnatural. Its PDF is the bell-shaped curve, familiar even to people who have not studied statistics.

The breadth of applicability of the Gaussian distribution follows in large part from a very powerful theoretical result, known as the *Central Limit Theorem*. Informally, the Central Limit Theorem states that in the limit, as the sample size becomes large, the sum (or, equivalently because it is proportional, the arithmetic mean) of a set of independent observations will have a Gaussian *sampling distribution*. That is, a histogram of the sums or sample means of a large number of different batches of the same kind of data, each of size n, will look like a bell curve if n is large enough. This is true regardless of the distribution from which the original data have been drawn. The data need not even be from the same distribution! Actually, the independence of the observations is not really necessary for the shape of the resulting distribution to be Gaussian either (see Section 5.2.4), which considerably broadens the applicability of the Central Limit Theorem for atmospheric data.

What is not clear for particular data sets is just how large the sample size must be for the Central Limit Theorem to apply. In practice this sample size depends on the distribution from which the summands are drawn. If the summed observations are themselves taken from a Gaussian distribution, the sum of any number of them (including, of course, $n = 1$) will also be Gaussian. For underlying distributions not too unlike the Gaussian (unimodal and not too asymmetrical), the sum of a modest number of observations will be nearly Gaussian. Summing daily temperatures to obtain a monthly averaged temperature is a good example of this situation. Daily temperature values can exhibit noticeable asymmetry (e.g., Figure 3.5), but are usually much more symmetrical than daily precipitation values. Conventionally, average daily temperature is approximated as the average of the daily maximum and minimum temperatures, so that the average monthly temperature is computed as

$$\overline{T} = \frac{1}{30}\sum_{i=1}^{30} \frac{T_{\max}(i) + T_{\min}(i)}{2}, \tag{4.23}$$

for a month with 30 days. Here the average monthly temperature is computed from the sum of 60 numbers drawn from two more or less symmetrical distributions. It is not surprising, in light of the Central Limit Theorem, that monthly temperature values are often very successfully represented by Gaussian distributions.

A contrasting situation is that of the monthly total precipitation, constructed as the sum of, say, 30 daily precipitation values. There are fewer numbers going into this sum than is the case for the average monthly temperature in Equation 4.23, but the more important difference has to do with the distribution of the underlying daily precipitation amounts. Typically most daily precipitation values are zero, and most of the nonzero amounts are small. That is, the distributions of daily precipitation amounts are usually very strongly skewed to the right (e.g., Figure 3.11b). Generally, the distribution of sums of 30 such values is also skewed to the right, although not so extremely. The schematic plot for $\lambda = 1$ in Figure 3.14 illustrates this asymmetry for total January precipitation at Ithaca. Note, however, that the distribution of Ithaca January precipitation totals in Figure 3.14 is much more symmetrical than the corresponding distribution for the underlying daily precipitation amounts in Figure 3.11b. Even though the summation of 30 daily values has not produced a Gaussian distribution for the monthly totals, the shape of the distribution of monthly precipitation is much closer to the Gaussian than the very strongly skewed distribution of the daily precipitation amounts. In humid climates, the distributions of seasonal (i.e., 90-day) precipitation totals begin to approach the Gaussian, but even annual precipitation totals at arid locations can exhibit substantial positive skewness. Whether the sample size is adequate for invocation of CLT for a particular data set can

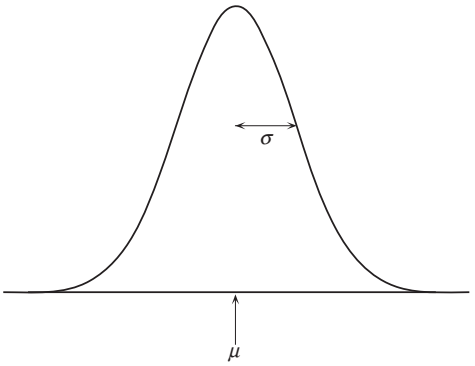

FIGURE 4.4 Probability density function for the Gaussian distribution, Equation 4.24. The mean, μ, locates the center of this symmetrical distribution, and the standard deviation, σ, controls the degree to which the distribution spreads out. Nearly all of the probability is within $\pm 3\sigma$ of the mean.

be assessed through bootstrapping (Section 5.3.5), examining the shape of the resulting bootstrap distribution for the sample mean.

The PDF for the Gaussian distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty. \tag{4.24}$$

The two distribution parameters are the mean, μ, and the standard deviation, σ; and π is the mathematical constant 3.14159…. Gaussian random variables are defined on the entire real line, so Equation 4.24 is valid for $-\infty < x < \infty$. Graphing Equation 4.24 results in the familiar bell-shaped curve shown in Figure 4.4. This figure shows that the mean locates the center of this symmetrical distribution, and the standard deviation controls the degree to which the distribution spreads out. Nearly all the probability is within $\pm 3\sigma$ of the mean.

In order to use the Gaussian distribution to represent a set of data, it is necessary to fit the two distribution parameters. Good parameter estimates for this distribution are easily obtained using the method of moments. Again, the method of moments amounts to nothing more than equating as many sample moments and distribution moments as there are parameters to be estimated. In the case of the Gaussian distribution the first two moments correspond exactly to the distribution parameters: the first moment is the mean, μ, and the second moment is the variance, σ^2. Therefore we simply estimate μ as the sample mean (Equation 3.2), and σ as the sample standard deviation (Equation 3.6).

If a data sample follows at least approximately a Gaussian distribution, these parameter estimates will make Equation 4.24 behave similarly to the data. Then, in principle, probabilities for events of interest can be obtained by integrating Equation 4.24. Practically, however, analytic integration of Equation 4.24 is impossible, so that a formula for the CDF, $F(x)$, for the Gaussian distribution does not exist. Rather, Gaussian probabilities are obtained in one of two ways. If the probabilities are needed as part of a computer program, the integral of Equation 4.24 can be economically approximated (e.g., Abramowitz and Stegun 1984) or computed by numerical integration (e.g., Press et al. 1986) to precision that is more than adequate. If only a few probabilities are needed, it is practical to compute them by hand using tabulated values such as those in Table B.1 in Appendix B.

In either of these two situations, a data transformation will nearly always be required. This is because Gaussian probability tables and algorithms pertain to the *standard Gaussian distribution*, that is, the Gaussian distribution having $\mu = 0$ and $\sigma = 1$. Conventionally, the random variable described by the standard Gaussian distribution is denoted as z. Its PDF simplifies to

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]. \tag{4.25}$$

The notation $\phi(z)$ is often used for the PDF of the standard Gaussian distribution, rather than $f(z)$. Similarly, $\Phi(z)$ is the conventional notation for the CDF of the standard Gaussian distribution, which is related to the "error function," erf(z) according to erf(z)=2Φ[$z\sqrt{(2)}$]-1. Any Gaussian random variable, x, can be transformed to standard form, z, simply by subtracting its mean and dividing by its standard deviation,

$$z = \frac{x - \mu}{\sigma}. \tag{4.26}$$

In practical settings, the mean and standard deviation usually need to be estimated using the corresponding sample statistics, so that we use

$$z = \frac{x - \bar{x}}{s}. \tag{4.27}$$

Note that whatever physical units characterize x will cancel in this transformation, so that the standardized variable, z, is always dimensionless.

Equation 4.27 is exactly the same as the standardized anomaly of Equation 3.27. Any batch of data can be transformed by subtracting the mean and dividing by the standard deviation, and this transformation will produce transformed values having a sample mean of zero and a sample standard deviation of one. However, the transformed data will not follow a Gaussian distribution unless the untransformed data do. Use of the standardized variable in Equation 4.27 to obtain Gaussian probabilities is illustrated in the following example.

Example 4.6. Evaluating Gaussian Probabilities

Consider a Gaussian distribution characterized by $\mu = 22.2°\text{F}$ and $\sigma = 4.4°\text{F}$. These parameters were fit to a set of average January temperatures at Ithaca. Suppose we are interested in evaluating the probability that an arbitrarily selected, or future, January will have average temperature as cold as or colder than 21.4°F, the value observed in 1987 (see Table A.1). Transforming this temperature using the standardization in Equation 4.26 yields $z = (21.4°\text{F} - 22.2°\text{F})/4.4°\text{F} = -0.18$. Thus the probability of a temperature as cold as or colder than 21.4°F is the same as the probability of a value of z as small as or smaller than -0.18: $\Pr\{X \le 21.4°\text{F}\} = \Pr\{Z \le -0.18\}$.

Evaluating $\Pr\{Z \le -0.18\}$ is easy, using Table B.1 in Appendix B, which contains cumulative probabilities for the standard Gaussian distribution, $\Phi(z)$. Looking across the row in Table B.1 labeled -0.1 to the column labeled 0.08 yields the desired probability, 0.4286. Evidently, there is a substantial probability that an average temperature this cold or colder will occur in January at Ithaca.

Notice that Table B.1 contains no rows for positive values of z. These are not necessary because the Gaussian distribution is symmetric. This means, for example, that $\Pr\{Z \ge +0.18\} = \Pr\{Z \le -0.18\}$, since there will be equal areas under the curve in Figure 4.4 to the left of $z = -0.18$, and to the right of $z = +0.18$. Therefore Table B.1 can be used more generally to evaluate probabilities for $z > 0$ by applying the relationship

$$\Pr\{Z \le z\} = 1 - \Pr\{Z \le -z\}, \tag{4.28a}$$

or, equivalently,

$$\Phi(z) = 1 - \Phi(-z), \tag{4.28b}$$

which follows from the fact that the total area under the curve of any PDF is 1 (Equation 4.18).

Using Equation 4.28 it is straightforward to evaluate $\Pr\{Z \le +0.18\} = 1 - 0.4286 = 0.5714$. The average January temperature at Ithaca to which $z = +0.18$ corresponds is obtained by inverting Equation 4.26,

$$x = \sigma z + \mu. \tag{4.29}$$

The probability is 0.5714 that an average January temperature at Ithaca will be no greater than (4.4°F)(0.18) + 22.2°F = 23.0°F.

It is only slightly more complicated to compute probabilities for outcomes between two specific values, say Ithaca January temperatures between 20°F and 25°F. Since the event $\{X \le 20°\text{F}\}$ is a subset of the event $\{X \le 25°\text{F}\}$, the desired probability, $\Pr\{20°\text{F} < T \le 25°\text{F}\}$ can be obtained by the subtraction $\Phi(z_{25}) - \Phi(z_{20})$. Here $z_{25} = (25.0°\text{F} - 22.2°\text{F})/4.4°\text{F} = 0.64$ and $z_{20} = (20.0°\text{F} - 22.2°\text{F})/4.4°\text{F} = -0.50$. Therefore (from Table B.1), $\Pr\{20°\text{F} < T \le 25°\text{F}\} = \Phi(z_{25}) - \Phi(z_{20}) = 0.739 - 0.309 = 0.430$.

It is also sometimes required to evaluate the inverse of the standard Gaussian CDF, that is, the standard Gaussian quantile function, $\Phi^{-1}(p)$. This function specifies values of the standard Gaussian variate, z, corresponding to particular cumulative probabilities, p. Again, an explicit formula for this function cannot be written, but Φ^{-1} can be evaluated using Table B.1 in reverse. For example, to find the average January Ithaca temperature defining the lowest decile (i.e., the coldest 10% of Januaries), the body of Table B.1 would be searched for $\Phi(z) = 0.10$. This cumulative probability corresponds almost exactly to $z = -1.28$. Using Equation 4.29, $z = -1.28$ corresponds to a January temperature of (4.4°F)(−1.28) + 22.2°F = 16.6°F. ◇

When high precision is not required for Gaussian probabilities, a "pretty good" approximation to the standard Gaussian CDF can be used,

$$\Phi(z) \approx \frac{1}{2}\left[1 \pm \sqrt{1 - \exp\left(\frac{-2z^2}{\pi}\right)}\right]. \tag{4.30}$$

The positive root is taken for $z > 0$ and the negative root is used for $z < 0$. The maximum errors produced by Equation 4.30 are about 0.003 (probability units) in magnitude, which occur at $z = \pm 1.65$. Equation 4.30 can be inverted to yield an approximation to the Gaussian quantile function, but the approximation is poor for the tail (i.e., for extreme) probabilities that are often of greatest interest.

Bivariate Normal Distribution

In addition to the power of the Central Limit Theorem, another reason that the Gaussian distribution is used so frequently is that it easily generalizes to higher dimensions. That is, it is usually straightforward to represent joint variations of multiple Gaussian variables through what is called the *multivariate Gaussian* or *multivariate normal distribution*. This distribution is discussed more extensively in Chapter 12, since in general the mathematical development for the multivariate Gaussian distribution requires use of matrix algebra.

However, the simplest case of the multivariate Gaussian distribution, describing the joint variations of two Gaussian variables, can be presented without vector notation. This two-variable distribution is known as the *bivariate Gaussian* or *bivariate normal distribution*. It is sometimes possible to use this distribution to describe the behavior of two non-Gaussian distributions if the variables are first subjected

to transformations such as those in Equation 3.21 or 3.24. In fact the opportunity to use the bivariate normal can be a major motivation for using such transformations.

Let the two variables considered be x and y. The bivariate normal distribution is defined by the joint PDF

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\}. \tag{4.31}$$

As a generalization of Equation 4.24 from one to two dimensions, this function defines a surface above the x–y plane rather than a curve above the x-axis. For continuous bivariate distributions, including the bivariate normal, probability corresponds geometrically to the volume under the surface defined by the PDF so that, analogously to Equation 4.18, necessary conditions to be fulfilled by any bivariate PDF are

$$\int\int_{x,y} f(x,y)\, dy\, dx = 1, \quad \text{and } f(x,y) \geq 0. \tag{4.32}$$

The bivariate normal distribution has five parameters: the two means and standard deviations for the variables x and y, and the correlation between them, ρ. The two marginal distributions for the variables x and y (i.e., the univariate PDFs $f(x)$ and $f(y)$) must both be Gaussian distributions and have parameters μ_x, σ_x, and μ_y, σ_y, respectively. It is usual, although not guaranteed, for the joint distribution of any two Gaussian variables to be bivariate normal. Fitting the bivariate normal distribution is very easy. The means and standard deviations are estimated using their sample counterparts for the x and y variables separately, and the parameter ρ is estimated as the Pearson product–moment correlation between x and y, Equation 3.28.

Figure 4.5 illustrates the general shape of the bivariate normal distribution. It is mound shaped in three dimensions, with properties that depend on the five parameters. The function achieves its maximum height above the point (μ_x, μ_y). Increasing σ_x stretches the density in the x direction and increasing σ_y stretches it in the y direction. For $\rho = 0$ the density is symmetric around the point (μ_x, μ_y) with respect to both the x- and y-axes. Curves of constant height (i.e., intersections of $f(x,y)$ with planes parallel to the x–y plane) are concentric circles if $\rho = 0$ and $\sigma_x = \sigma_y$, and are concentric ellipses otherwise. As ρ increases in absolute value the density function is stretched diagonally, with the curves of constant height becoming increasingly elongated ellipses. For negative ρ the orientation of these

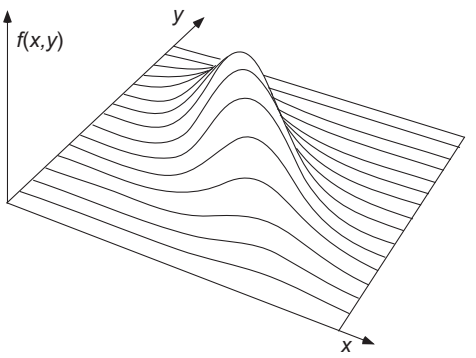

FIGURE 4.5 Perspective view of a bivariate normal distribution with $\sigma_x = \sigma_y$, and $\rho = -0.75$. The individual lines depicting the hump of the bivariate distribution have the shapes of (univariate) Gaussian distributions, illustrating that conditional distributions of x given a particular value of y are themselves Gaussian.

ellipses is as depicted in Figure 4.5: larger values of x are more likely to occur simultaneously with smaller values of y, and smaller values of x are more likely with larger values of y. The ellipses have the opposite orientation (positive slope) for positive values of ρ.

Probabilities for joint outcomes of x and y are given by the double integral of Equation 4.31 over the relevant region in the plane, for example

$$\Pr\{(y_1 < Y \le y_2) \cap (x_1 < X \le x_2)\} = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)\, dy\, dx. \tag{4.33}$$

This integration cannot be done analytically, and in practice numerical methods usually are used. Probability tables for the bivariate normal distribution do exist (National Bureau of Standards 1959), but they are lengthy and cumbersome. It is possible to compute probabilities for elliptically shaped regions, called probability ellipses, centered on (μ_x, μ_y) using the method illustrated in Example 12.1. When computing probabilities for other regions, it can be more convenient to work with the bivariate normal distribution in standardized form. This is the extension of the standardized univariate Gaussian distribution (Equation 4.25) and is achieved by subjecting both the x and y variables to the transformation in Equation 4.26 or 4.27. Thus $\mu_{z_x} = \mu_{z_y} = 0$ and $\sigma_{z_x} = \sigma_{z_y} = 1$, leading to the bivariate density

$$\phi\left(z_x, z_y\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{z_x^2 - 2\rho\, z_x z_y + z_y^2}{2\left(1-\rho^2\right)}\right]. \tag{4.34}$$

A very useful property of the bivariate normal distribution is that the conditional distribution of one of the variables, given any particular value of the other, is univariate Gaussian. This property is illustrated graphically in Figure 4.5, where the individual lines defining the shape of the distribution in three dimensions themselves have Gaussian shapes. Each indicates a function proportional to a conditional distribution of x given a particular value of y. The parameters for these conditional Gaussian distributions can be calculated from the five parameters of the bivariate normal distribution. For the conditional distribution of x given a particular value of y, the conditional Gaussian density function $f(x|Y = y)$ has parameters

$$\mu_{x|y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}\left(y - \mu_y\right) \tag{4.35a}$$

and

$$\sigma_{x|y} = \sigma_x\sqrt{1-\rho^2}. \tag{4.35b}$$

Equation 4.35a relates the mean of x to the distance of y from its mean, scaled according to the product of the correlation and the ratio of the standard deviations. It indicates that the conditional mean $\mu_{x|y}$ is larger than the unconditional mean μ_x if y is greater than its mean and ρ is positive, or if y is less than its mean and ρ is negative. If x and y are uncorrelated, knowing a value of y gives no additional information about y, and $\mu_{x|y} = \mu_x$ since $\rho = 0$. Equation 4.35b indicates that, unless the two variables are uncorrelated, $\sigma_{x|y} < \sigma_x$, regardless of the sign of ρ. Knowing y provides some information about x if the two have nonzero correlation, and the diminished uncertainty about x is reflected by the smaller standard deviation. In this sense, ρ^2 is often interpreted as the proportion of the variance in x that is accounted for by y.

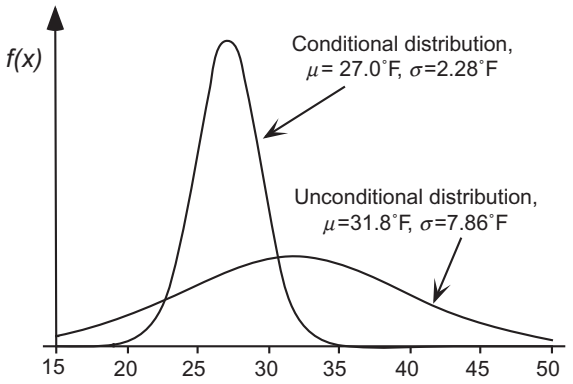

FIGURE 4.6 Gaussian distributions representing the unconditional distribution for daily January maximum temperature at Canandaigua, and the conditional distribution given that the Ithaca maximum temperature was 25°F. The large correlation between maximum temperatures at the two locations results in the conditional distribution being much sharper, reflecting substantially diminished uncertainty.

Example 4.7. Bivariate Normal Distribution and Conditional Probability

Consider the maximum temperature data for January 1987 at Ithaca and Canandaigua, in Table A.1. Figure 3.5 indicates that these data are fairly symmetrical, so that it may be reasonable to model their joint behavior as bivariate normal. A scatterplot of these two variables is shown in one of the panels of Figure 3.31. The average maximum temperatures are 29.87°F and 31.77°F at Ithaca and Canandaigua, respectively. The corresponding sample standard deviations are 7.71°F and 7.86°F. Table 3.5 indicates their Pearson correlation to be 0.957.

With such a large correlation, knowing the temperature at one location should give very strong information about the temperature at the other. Suppose it is known that the Ithaca maximum temperature is 25°F, and probability information about the Canandaigua maximum temperature is needed, perhaps for the purpose of estimating a missing value there. Using Equation 4.35a, the conditional mean for the distribution of maximum temperature at Canandaigua, given that the Ithaca maximum temperature is 25°F, is 27.0°F—substantially lower than the unconditional mean of 31.77°F. Using Equation 4.35b, the conditional standard deviation is 2.28°F. This would be the conditional standard deviation regardless of the particular value of the Ithaca temperature chosen, since Equation 4.35b does not depend on the value of the conditioning variable. The conditional standard deviation is so much lower than the unconditional standard deviation because of the high correlation of maximum temperatures between the two locations. As illustrated in Figure 4.6, this reduced uncertainty means that any of the conditional distributions for Canandaigua temperature given the Ithaca temperature will be much sharper than the unmodified, unconditional distribution for Canandaigua maximum temperature.

Using these parameters for the conditional distribution of maximum temperature at Canandaigua, we can compute such quantities as the probability that the Canandaigua maximum temperature is at or below freezing, given that the Ithaca maximum is 25°F. The required standardized variable is $z = (32 - 27.0)/2.28 = 2.19$, which corresponds to a probability of 0.986. By contrast, the corresponding climatological probability (without benefit of knowing the Ithaca maximum temperature) would be computed from $z = (32 - 31.8)/7.86 = 0.025$, corresponding to the much lower probability 0.510. ◇

4.4.3. Some Gaussian Distribution Variants

Although it is mathematically possible to fit Gaussian distributions to data that are nonnegative and positively skewed, the results are generally not useful. For example, the January 1933–1982 Ithaca

precipitation data in Table A.2 can be characterized by a sample mean of 1.96 in. and a sample standard deviation of 1.12 in. These two statistics are sufficient to fit a Gaussian distribution to these data, and this distribution is shown as the dashed PDF in Figure 4.16, but applying this fitted distribution leads to nonsense. In particular, using Table B.1, we can compute the probability of negative precipitation as $\Pr\{Z < (0.00 - 1.96)/1.12\} = \Pr\{Z < -1.75\} = 0.040$. This computed probability is not especially large, but neither is it vanishingly small. The true probability is exactly zero: observing negative precipitation is impossible.

Several variants of the Gaussian distribution exist that can be useful for representing these kinds of data. Three of these are presented in this section: the lognormal distribution, the truncated Gaussian distribution, and the censored Gaussian distribution.

Lognormal Distribution

As noted in Section 3.4.1, one approach to dealing with skewed data is to subject them to a power transformation that produces an approximately Gaussian distribution. When that power transformation is logarithmic (i.e., $\lambda = 0$ in Equation 3.21), the (original, untransformed) data are said to follow the *lognormal distribution*, with PDF

$$f(x) = \frac{1}{x\,\sigma_y\sqrt{2\pi}} \exp\left[-\frac{\left(\ln x - \mu_y\right)^2}{2\sigma_y^2}\right], \quad x > 0. \tag{4.36}$$

Here μ_y and σ_y are the mean and standard deviation, respectively, of the transformed variable, $y = \ln(x)$. Actually, the lognormal distribution is somewhat confusingly named, since the random variable x is the *anti*log of a variable y that follows a Gaussian distribution.

Parameter fitting for the lognormal distribution is simple and straightforward: the mean and standard deviation of the log-transformed data values y—that is, μ_y and σ_y, respectively—are estimated by their sample counterparts. The relationships between these parameters, in Equation 4.36, and the mean and variance of the original variable X are

$$\mu_x = \exp\left[\mu_y + \frac{\sigma_y^2}{2}\right] \tag{4.37a}$$

and

$$\sigma_x^2 = \left(\exp\left[\sigma_y^2\right] - 1\right) \exp\left[2\mu_y + \sigma_y^2\right]. \tag{4.37b}$$

Lognormal probabilities are evaluated simply by working with the transformed variable $y = \ln(x)$, and using computational routines or probability tables for the Gaussian distribution. In this case the standard Gaussian variable

$$z = \frac{\ln(x) - \mu_y}{\sigma_y}, \tag{4.38}$$

follows a Gaussian distribution with $\mu_z = 0$ and $\sigma_z = 1$.

The lognormal distribution is sometimes somewhat arbitrarily assumed for positively skewed data. In particular, the lognormal too frequently is used without checking whether a different power

transformation might produce more nearly Gaussian behavior. In general, it is recommended that other candidate power transformations also be investigated as explained in Section 3.4.1 before the lognormal distribution is adopted to represent a particular data set.

Truncated Gaussian Distribution

The *truncated Gaussian distribution* provides another approach to modeling positive and positively skewed data using the Gaussian bell-curve shape. Although it is mathematically possible to fit a conventional Gaussian distribution to such data in the usual way, unless both the mean is positive and the standard deviation is much smaller than the mean, the resulting Gaussian distribution will specify substantial probability for impossible negative outcomes. A zero-truncated Gaussian distribution is a Gaussian distribution having nonzero probability only for positive values of the random variable, with the portion of the PDF corresponding to negative values cut off, or truncated. Zero-truncated Gaussian distributions have been used successfully to represent distributions of forecast wind speeds (e.g., Thorarinsdottir and Gneiting 2010, Baran 2014).

Figure 4.3 illustrates a typical zero-truncated Gaussian shape, for which $\mu > 0$ and σ is comparable in magnitude to μ. The truncated Gaussian PDF is written in terms of the PDF and CDF of the standard Gaussian distribution, and has the form

$$f(x) = \phi\left(\frac{x-\mu}{\sigma}\right)\left[\sigma\,\Phi\left(\frac{\mu}{\sigma}\right)\right]^{-1}, \quad x > 0. \tag{4.39}$$

The factor $\Phi(\mu/\sigma)$ in the denominator is necessary in order that Equation 4.39 will integrate to unity. In effect, the probability for $x \le 0$ has been spread proportionally across the rest of the distribution. The corresponding CDF is then

$$F(x) = \left[\Phi\left(\frac{x-\mu}{\sigma}\right) - \Phi\left(\frac{-\mu}{\sigma}\right)\right] \Big/ \Phi\left(\frac{\mu}{\sigma}\right), \tag{4.40}$$

which converges to $\Phi[(x-\mu)/\sigma]$ (i.e., an ordinary Gaussian CDF) for $\mu \gg \sigma$.

Estimating the distribution location parameter μ and scale parameter σ is best done using iterative fitting methods (e.g., Thorarinsdottir and Gneiting 2010) such as maximum likelihood (Section 4.6). Because Equation 4.39 specifies zero probability for $x \le 0$, the mean of a zero-truncated Gaussian distribution is necessarily larger than the location parameter μ, and the variance is necessarily smaller than the scale parameter σ^2. Specifically,

$$E[x] = \mu + \sigma\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)} \tag{4.41a}$$

and

$$Var[x] = \sigma^2\left\{1 - \frac{\mu}{\sigma}\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)} - \left[\frac{\phi(\mu/\sigma)}{\Phi(\mu/\sigma)}\right]^2\right\}. \tag{4.41b}$$

Analogously to Equation 4.40, Equations 4.41a and 4.41b converge to μ and σ^2, respectively, for $\mu \gg \sigma$.

Censored Gaussian Distribution

The *censored Gaussian distribution* is an alternative to truncation that may be more appropriate for representing precipitation data, where there is a discrete probability mass at exactly zero. Zero censoring, in contrast to zero truncation, allows a probability distribution to represent values falling hypothetically

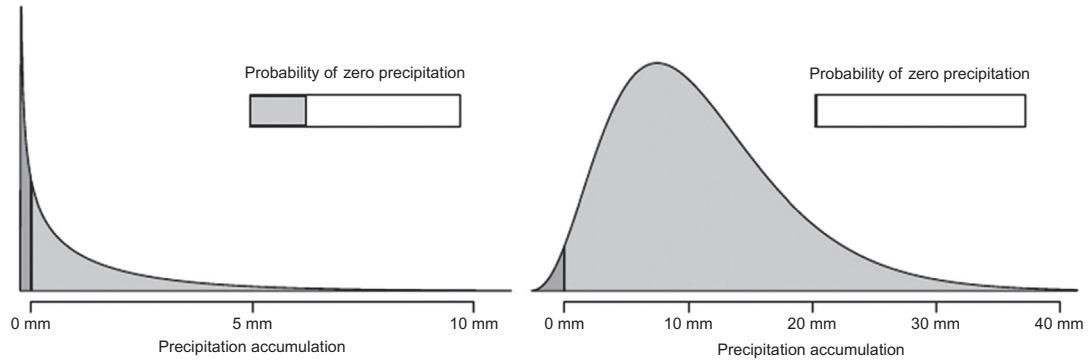

FIGURE 4.7 Illustration of the use of censoring to represent the discontinuous probability spike for zero precipitation, based on shifted-gamma (Pearson III) distributions. Probability for negative amounts is attributed to exactly zero precipitation, as indicated by the inset bars. *From Scheuerer and Hamill (2015a). © American Meteorological Society. Used with permission.*

below a censoring threshold, even though those values have not been observed. The censoring threshold is generally zero for precipitation data, and any probability corresponding to negative values is assigned to exactly zero, yielding a probability spike there. The idea is distinct from truncation, where probability for any negative values is spread proportionally across the positive values.

Bárdossy and Plate (1992) used zero-censored Gaussian distributions to represent (power-transformed) precipitation data, enabling a spatial precipitation model using the multivariate Gaussian distribution (Chapter 12). The censored distribution idea may also be applied to distributions other than the Gaussian in order to represent the discontinuous probability spike at zero precipitation, including the generalized extreme-value distribution (Section 4.4.7) (Scheuerer 2014), the gamma distribution (Section 4.4.5) (Wilks 1990), the shifted-gamma (Pearson III, Section 4.4.5) (Scheuerer and Hamill 2015a, Baran and Nemoda 2016), and the logistic distribution (Section 4.4.4) (Stauffer et al. 2017). Figure 4.7 illustrates the idea for two censored distributions. Probability for nonzero precipitation amounts is distributed according to the fitted distributions. Any probability that those distributions attribute to negative amounts is assigned to exactly zero precipitation, as indicated by the shaded inset bars.

4.4.4. Logistic Distributions

The *logistic distribution* shape is quite similar to that of the Gaussian distribution, but with somewhat heavier tails. Its two parameters are the location, μ, and scale, σ. The PDF and CDF for the logistic distribution are, respectively,

$$f(x) = \exp\left(\frac{x-\mu}{\sigma}\right)\left(1+\exp\left(\frac{x-\mu}{\sigma}\right)\right)^{-2} \tag{4.42}$$

and

$$F(x) = \frac{\exp\left(\frac{x-\mu}{\sigma}\right)}{1+\exp\left(\frac{x-\mu}{\sigma}\right)}. \tag{4.43}$$

The mean of the logistic distribution is equal to the location parameter μ, but the variance is $\sigma^2\pi^2/3$, so that logistic distributions with unit variance have $\sigma = \sqrt{3}/\pi$. Figure 4.8 compares the standard (zero mean and unit variance) logistic distribution (solid curves) and the corresponding standard Gaussian distribution

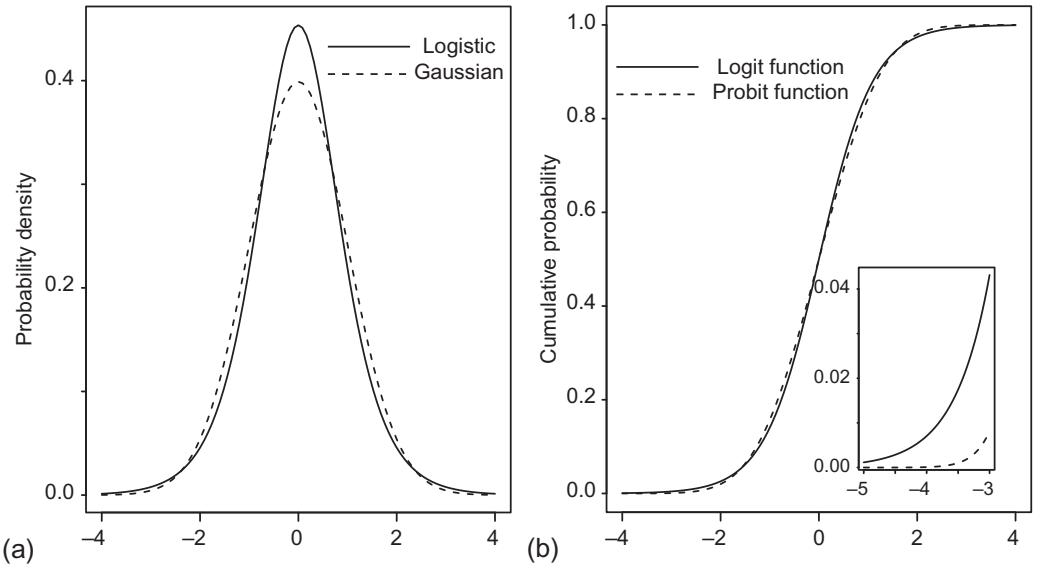

FIGURE 4.8 Comparison of logistic *(solid)* and Gaussian *(dashed)* probability densities (a), and CDFs (b). Inset in panel (b) shows detail of left-tail behavior. Both distributions shown have zero mean and unit variance.

(dashed curves), in terms of (a) their PDFs and (b) CDFs. As indicated in the inset in Figure 4.8b, the heavier tails of the logistic distribution are most evident for more extreme values. For example, $\Pr\{x \le -4\}$ is 0.00003 for the standard Gaussian distribution (cf. Table B.1), but 0.0007 for the logistic distribution according to Equation 4.43. The CDF for the logistic distribution is called the *logit function*, and the very similar CDF for standard Gaussian distribution is called the *probit function*.

Since the logistic CDF can be written in closed form, so also can its quantile function, which is

$$F^{-1}(p) = \mu + \sigma \ln\left(\frac{p}{1-p}\right). \tag{4.44}$$

For $\mu = 0$ and $\sigma = 1$, Equation 4.44 is equivalent to the log-odds transformation (Equation 3.24).

4.4.5. Gamma Distributions

The statistical distributions of many atmospheric variables are distinctly asymmetric and skewed to the right. Often the skewness occurs when there is a physical limit on the left that is relatively near the range of the data. Common examples are precipitation amounts or wind speeds, which are physically constrained to be nonnegative. The *gamma distribution* is a common choice for representing such data.

The gamma distribution is defined by the PDF

$$f(x) = \frac{(x/\beta)^{\alpha-1}\exp(-x/\beta)}{\beta\,\Gamma(\alpha)}\ ,\ x,\alpha,\beta > 0. \tag{4.45}$$

The two parameters of the distribution are α, the shape parameter; and β, the scale parameter. The quantity $\Gamma(\alpha)$ is the gamma function, defined in Equation 4.7, evaluated at α. The PDF of the gamma

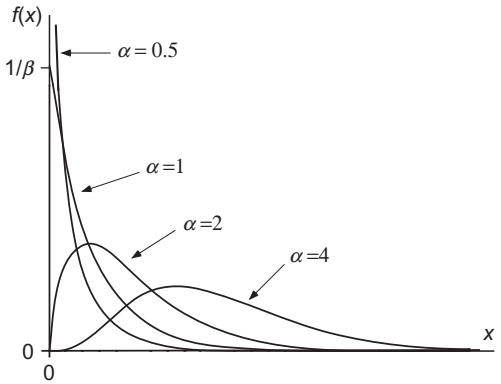

FIGURE 4.9 Gamma distribution density functions for four values of the shape parameter, α.

distribution takes on a wide range of shapes depending on the value of the shape parameter, α. As illustrated in Figure 4.9, for $\alpha < 1$ the distribution is very strongly skewed to the right, with $f(x) \rightarrow \infty$ as $x \rightarrow 0$. For $\alpha = 1$ the function intersects the vertical axis at $1/\beta$ for $x = 0$ (this special case of the gamma distribution is called the *exponential distribution*, which is described more fully later in this section). For $\alpha > 1$ the gamma distribution density function begins at the origin, $f(0) = 0$. Progressively larger values of α result in less skewness, and a shifting of probability density to the right. For very large values of α (larger than perhaps 50–100) the gamma distribution approaches the Gaussian distribution in form.

The parameter α is always dimensionless. The role of the scale parameter, β, effectively is to stretch or squeeze (i.e., to scale) the gamma density function to the right or left, depending on the overall magnitudes of the data values represented. Notice that the random quantity x in Equation 4.45 is divided by β in both places where it appears. The scale parameter β has the same physical dimensions as x. As the distribution is stretched to the right by larger values of β, its height must drop in order to satisfy Equation 4.18, and conversely as the density is squeezed to the left its height must rise. These adjustments in height are accomplished by the β in the denominator of Equation 4.45. The versatility in shape of the gamma distribution makes it an attractive candidate for representing precipitation data. It is often used for this purpose, although observed precipitation data often exhibit heavier tails than corresponding fitted gamma distributions, especially at smaller spatial scales and for shorter accumulation periods (e.g., Cavanaugh and Gershunov 2015, Katz et al., 2002, Wilks 1999a).

The gamma distribution is more difficult to work with than the Gaussian distribution because obtaining good parameter estimates from particular batches of data is not as straightforward. The simplest (although certainly not best) approach to fitting a gamma distribution is to use the method of moments. Even here, however, there is a complication, because the two parameters for the gamma distribution do not correspond exactly to moments of the distribution, as is the case for the Gaussian distribution. The mean of the gamma distribution is given by the product $\alpha\beta$, and the variance is $\alpha\beta^2$. Equating these expressions with the corresponding sample quantities yields a set of two equations in two unknowns, which can be solved to yield the moments estimators

$$\hat{\alpha} = \overline{x}^2 / s^2 \tag{4.46a}$$

and

$$\hat{\beta} = s^2 / \overline{x}. \tag{4.46b}$$

The moments estimators for the gamma distribution are usually reasonably accurate for large values of the shape parameter, perhaps $\alpha > 10$, but can yield poor results for small values of α (Thom 1958, Wilks 1990). The moments estimators in this case are said to be inefficient, in the technical sense of not making maximum use of the information in a data set. The practical consequence of this inefficiency is that particular values of the parameters calculated using Equation 4.46 are erratic, or unnecessarily variable, from data sample to data sample.

A much better approach to parameter fitting for the gamma distribution is to use the method of *maximum likelihood*. For many distributions, including the gamma distribution, maximum likelihood fitting requires an iterative procedure that is really only practical using a computer. Section 4.6 presents the method of maximum likelihood for fitting parametric distributions, including the gamma distribution in Example 4.13.

Approximations to the *maximum likelihood estimators* (MLEs) for the gamma distribution are available that do not require iterative computation. Two of these employ the sample statistic

$$D = \ln(\bar{x}) - \frac{1}{n}\sum_{i=1}^{n} \ln(x_i), \tag{4.47}$$

which is the difference between the natural log of the sample mean and the mean of the logs of the data. Equivalently, the sample statistic D is the difference between the logs of the arithmetic and geometric means. Notice that the sample mean and standard deviation are not sufficient to compute the statistic D, since each data value must be used to compute the second term in Equation 4.47.

A simple approximation to maximum likelihood estimation for the gamma distribution is due to Thom (1958). The Thom estimator for the shape parameter is

$$\hat{\alpha} = \frac{1 + \sqrt{1 + 4D/3}}{4D}, \tag{4.48}$$

after which the scale parameter is obtained from

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}}. \tag{4.49}$$

An alternative approach is a polynomial approximation to the shape parameter (Greenwood and Durand 1960). One of two equations is used,

$$\hat{\alpha} = \frac{0.5000876 + 0.1648852D - 0.0544274D^2}{D}, \quad 0 \le D \le 0.5772, \tag{4.50a}$$

or

$$\hat{\alpha} = \frac{8.898919 + 9.059950D + 0.9775373D^2}{17.79728D + 11.968477D^2 + D^3}, \quad 0.5772 \le D \le 17.0, \tag{4.50b}$$

depending on the value of D. The scale parameter is again subsequently estimated using Equation 4.49.

Another approximation to maximum likelihood for the gamma distribution is presented in Ye and Chen (2017). Because all three of these approximations require computing logarithms of the data, none are suitable if some values are exactly zero. In that case a workaround using the censoring concept can be used (Wilks 1990).

As was the case for the Gaussian distribution, the gamma density function is not analytically integrable. Gamma distribution probabilities must therefore be obtained either by computing

approximations to the CDF (by numerically integrating Equation 4.45) or from tabulated probabilities. Formulas and computer routines for this purpose can be found in Abramowitz and Stegun (1984) and Press et al. (1986), respectively. A table of gamma distribution probabilities is included as Table B.2 in Appendix B.

In any of these cases, gamma distribution probabilities will be available for the *standard gamma distribution*, with $\beta = 1$. Therefore it is nearly always necessary to transform by rescaling the variable X of interest (characterized by a gamma distribution with arbitrary scale parameter β) to the standardized variable

$$\xi = x/\beta, \tag{4.51}$$

which follows a gamma distribution with $\beta = 1$. The standard gamma variate ξ is dimensionless. The shape parameter, α, will be the same for both x and ξ. The procedure is analogous to the transformation to the standardized Gaussian variable, z, in Equations 4.26 and 4.27.

Cumulative probabilities for the standard gamma distribution are given by a mathematical function known as the *incomplete gamma function*, $P(\alpha, \xi) = \Pr\{\Xi \le \xi\} = F(\xi)$. It is this function that was used to compute the probabilities in Table B.2. The cumulative probabilities for the standard gamma distribution in Table B.2 are arranged in an inverse sense to the Gaussian probabilities in Table B.1. That is, quantiles (transformed data values, ξ) of the distributions are presented in the body of the table, and cumulative probabilities are listed as the column headings. Different probabilities are obtained for different shape parameters, α, which appear in the first column.

Example 4.8. Evaluating Gamma Distribution Probabilities

Consider the data for January precipitation at Ithaca during the 50 years 1933–1982 in Table A.2. The average January precipitation for this period is 1.96 in. and the mean of the logarithms of the monthly precipitation totals is 0.5346, so Equation 4.47 yields $D = 0.139$. Both Thom's method (Equation 4.48) and the Greenwood and Durand formula (Equation 4.50a) yield $\alpha = 3.76$ and $\beta = 0.52$ in. This result agrees well with the maximum likelihood values obtained using the method outlined in Example 4.13. In contrast, the moments estimators (Equation 4.46) yield $\alpha = 3.09$ and $\beta = 0.64$ in.

Adopting the approximate MLEs, the unusualness of the January 1987 precipitation total at Ithaca can be evaluated with the aid of Table B.2. That is, by representing the climatological variations in Ithaca January precipitation by the fitted gamma distribution with $\alpha = 3.76$ and $\beta = 0.52$ in., the cumulative probability corresponding to 3.15 in. (the sum of the daily values for Ithaca in Table A.1) can be computed.

First, applying Equation 4.51, the standard gamma variate $\xi = 3.15$ in./0.52 in. = 6.06. Adopting $\alpha = 3.75$ as the closest tabulated value to the fitted $\alpha = 3.76$, it can be seen that $\xi = 6.06$ lies between the tabulated values for $F(5.214) = 0.80$ and $F(6.354) = 0.90$. Interpolation yields $F(6.06) = 0.874$, indicating that there is approximately one chance in eight for a January this wet or wetter to occur at Ithaca. The probability estimate could be refined slightly by interpolating between the rows for $\alpha = 3.75$ and $\alpha = 3.80$ to yield $F(6.06) = 0.873$, although this additional calculation would probably not be worth the effort.

Table B.2 can also be used to invert the gamma CDF to find precipitation values corresponding to particular cumulative probabilities, $\xi = F^{-1}(p)$, that is, to evaluate the quantile function. Dimensional precipitation values are then recovered by reversing the transformation in Equation 4.51. Consider estimation of the median January precipitation at Ithaca. This will correspond to the value of ξ satisfying $F(\xi) = 0.50$ which, in the row for $\alpha = 3.75$ in Table B.2, is 3.425. The corresponding dimensional precipitation amount is given by the product $\xi\beta = (3.425)(0.52\text{ in.}) = 1.78$ in. By comparison, the sample median

of the precipitation data in Table A.2 is 1.72 in. It is not surprising that the median is less than the mean of 1.96 in., since the distribution is positively skewed. A (perhaps surprising, but often unappreciated) fact exemplified by this comparison is that below "normal" (i.e., below average) precipitation is typically more likely than above normal precipitation, as a consequence of the positive skewness of the distribution of precipitation. ◇

Example 4.9. Gamma Distribution in Operational Climatology, I. Reporting Seasonal Outcomes
The gamma distribution can be used to report monthly and seasonal precipitation amounts in a way that allows comparison with locally applicable climatological distributions. Figure 4.10 shows an example of this format for global precipitation for March–May 2014. The precipitation amounts for this 3-month period are not shown as accumulated depths, but rather as quantiles corresponding to local climatological gamma distributions. Five categories are mapped: less than the 10th percentile $q_{0.1}$, between the 10th and 30th percentile $q_{0.3}$, between the 30th and 70th percentile $q_{0.7}$, between the 70th and 90th percentile $q_{0.9}$, and wetter than the 90th percentile.

It is immediately clear which regions received substantially less, slightly less, about the same, slightly more, or substantially more precipitation during this period as compared to the underlying climatological distributions, even though the shapes of these distributions vary widely. One of the advantages of expressing precipitation amounts in terms of climatological gamma distributions is that these very strong differences in the shapes of the precipitation climatologies do not confuse comparisons between locations. Figure 4.11 illustrates the definition of the percentiles using a gamma PDF with $\alpha = 2$. The distribution is divided into five categories corresponding to the five shading levels in Figure 4.10, with the precipitation amounts $q_{0.1}$, $q_{0.3}$, $q_{0.7}$, and $q_{0.9}$ separating regions of the distribution containing 10%, 20%, 40%, 20%, and 10% of the probability, respectively. ◇

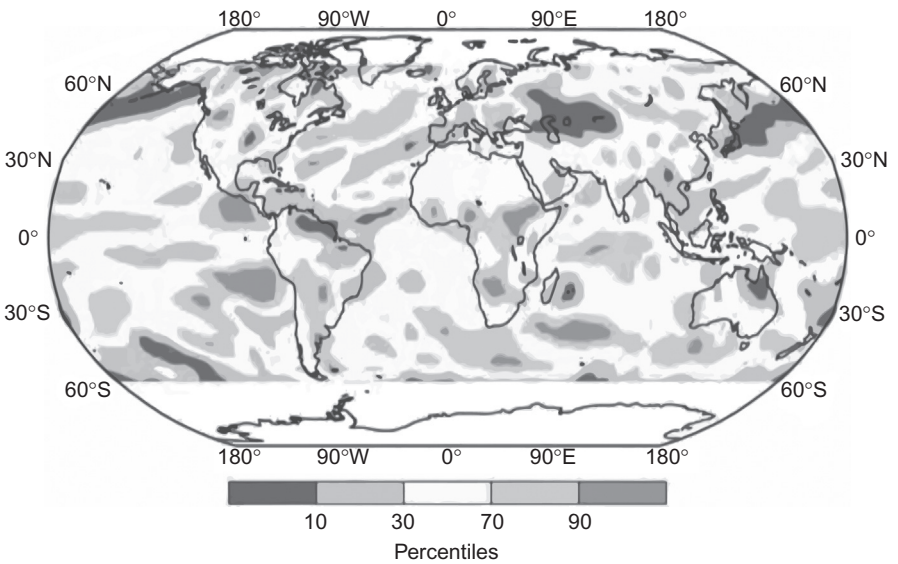

FIGURE 4.10 Global precipitation totals for the period March–May 2014, expressed as percentile values of local gamma distributions. *From Blunden and Arndt (2015). © American Meteorological Society. Used with permission.*

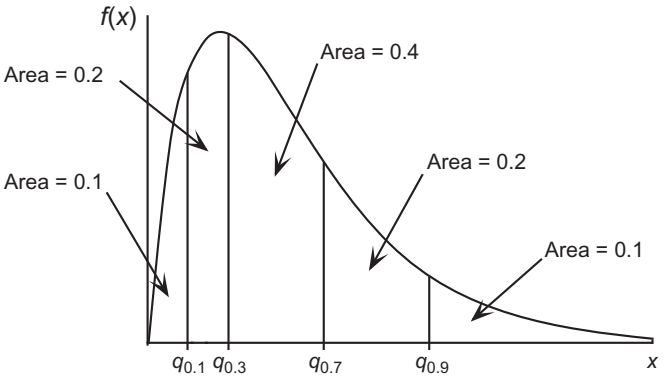

FIGURE 4.11 Illustration of the precipitation categories in Figure 4.10 in terms of a gamma distribution density function with $\alpha = 2$. Outcomes drier than the 10th percentile lie to the left of $q_{0.1}$. Areas with precipitation between the 30th and 70th percentiles (between $q_{0.3}$ and $q_{0.7}$) would be unshaded on the map. Precipitation in the wettest 10% of the climatological distribution lies to the right of $q_{0.9}$.

Exponential Distribution

There are several important special cases of the gamma distribution, which result from particular restrictions on the parameters α and β. For $\alpha = 1$, the gamma distribution reduces to the *exponential distribution*, with PDF

$$f(x) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right), \quad x \geq 0. \tag{4.52}$$

The shape of this density is simply an exponential decay, as indicated in Figure 4.7, for $\alpha = 1$. Equation 4.52 is analytically integrable, so the CDF for the exponential distribution exists in closed form,

$$F(x) = 1 - \exp\left(-\frac{x}{\beta}\right). \tag{4.53}$$

The quantile function is easily derived by solving Equation 4.53 for x (Equation 4.94). Since the shape of the exponential distribution is fixed by the restriction $\alpha = 1$, it is usually not suitable for representing variations in quantities like precipitation, although mixtures of two exponential distributions (see Section 4.4.9) can represent daily nonzero precipitation values quite well.

An important use of the exponential distribution in atmospheric science is in the characterization of the size distribution of raindrops, called drop-size distributions (e.g., Sauvageot 1994). When the exponential distribution is used for this purpose, it is called the *Marshall–Palmer distribution*, and generally denoted $N(D)$, which indicates a distribution over the numbers of droplets as a function of their diameters. Drop-size distributions are particularly important in radar applications where, for example, reflectivities are computed as expected values of a quantity called the backscattering cross-section, with respect to a drop-size distribution such as the exponential.

Erlang Distribution

The second special case of the gamma distribution is the *Erlang distribution*, in which the shape parameter α is restricted to integer values. One application of the Erlang distribution is as the distribution of waiting times until the αth Poisson event, for the Poisson rate $\mu = 1/\beta$.

Chi-Square Distribution

The *chi-square* (χ^2) distribution is a yet another special case of the gamma distribution. Chi-square distributions are gamma distributions with scale parameter $\beta = 2$. Chi-square distributions are conventionally written in terms of an integer-valued parameter called the *degrees of freedom*, denoted ν. The relationship to the gamma distribution more generally is that the degrees of freedom are twice the gamma distribution shape parameter, or $\alpha = \nu/2$, yielding the chi-square PDF

$$f(x) = \frac{x^{(\nu/2-1)} \exp(-x/2)}{2^{\nu/2}\,\Gamma(\nu/2)}, \quad x > 0. \tag{4.54}$$

Since it is the gamma scale parameter that is fixed at $\beta = 2$ to define the chi-square distribution, Equation 4.54 is capable of the same variety of shapes as the full gamma distribution, as the shape parameter α varies. Because there is no explicit horizontal scale in Figure 4.9, it could be interpreted as showing chi-square densities with $\nu = 1$, 2, 4, and 8 degrees of freedom. The chi-square distribution arises as the distribution of the sum of ν squared independent standard Gaussian variates and is used in several ways in the context of statistical inference (see Chapters 5 and 12). Table B.3 lists right-tail quantiles for chi-square distributions.

Pearson III Distribution

The gamma distribution is also sometimes generalized to a three-parameter distribution by moving the PDF to the left or right according to a shift parameter ζ. This three-parameter gamma distribution is also known as the Pearson Type III, or simply *Pearson III distribution*, and has PDF

$$f(x) = \frac{\left(\dfrac{x-\zeta}{\beta}\right)^{\alpha-1} \exp\left(-\dfrac{x-\zeta}{\beta}\right)}{|\beta\,\Gamma(\alpha)|}, \quad x > \zeta \text{ for } \beta > 0, \text{ or } x < \zeta \text{ for } \beta < 0. \tag{4.55}$$

This distribution is also sometimes called the *shifted gamma distribution*. Usually the scale parameter β is positive, which results in the Pearson III being a gamma distribution shifted to the right if $\zeta > 0$, with support $x > \zeta$. However, Equation 4.55 also allows $\beta < 0$, in which case the PDF is reflected (and so has a long left tail and negative skewness), and the support is $x < \zeta$.

Generalized Gamma Distribution

Sometimes, analogously to the lognormal distribution, the random variable x in Equation 4.55 has been log-transformed, in which case the distribution of the original variable [=exp(x)] is said to follow the *log-Pearson III distribution*, which is commonly used in hydrologic applications (e.g., Griffis and Stedinger 2007). Other power transformations besides the logarithm can also be used, and in that case the resulting distributions typically are formulated without the shift parameter ζ in Equation 4.55, yielding the generalized gamma distribution (Stacy 1962), with PDF

$$f(x) = \frac{\lambda x^{\alpha\lambda-1}}{\beta^{\alpha\lambda}\Gamma(\alpha)} \exp\left[-(x/\beta)^{\lambda}\right], \quad x > 0. \tag{4.56}$$

Here the power transformation parameter λ functions in the same way as in Equation 3.20.

Example 4.10. Gamma Distribution in Operational Climatology, II. The Standardized Precipitation Index

The *Standardized Precipitation Index* (SPI) is a popular approach to characterizing drought or wet-spell conditions, by expressing precipitation for monthly or longer periods in terms of the corresponding climatological distribution. McKee et al. (1993) originally proposed using gamma distributions for this purpose, and Guttman (1999) has suggested using the Pearson III distribution.

Computation of the SPI is accomplished through the *normal quantile transform*,

$$z = \Phi^{-1}[F(x)], \tag{4.57}$$

which is an instance of the operation sometimes also referred to as *quantile mapping*. The SPI is equal to z in Equation 4.57, so that a precipitation value x is characterized in terms of the standard Gaussian variate z that yields the same cumulative probability as the original data that follows the distribution $F(x)$. The SPI is thus a probability index that expresses precipitation deficits (SPI < 0) or excesses (SPI > 0) in a standardized way, accounting for differences in precipitation climatologies due to geographic and/or timescale differences. Precipitation accumulations characterized by $|\text{SPI}| > 1.0$, >1.5, and >2.0 are qualitatively and somewhat arbitrarily characterized as being dry or wet, moderately dry or wet, and extremely dry or wet, respectively (Guttman 1999).

Consider computing the SPI for the January 1987 Ithaca precipitation accumulation of 3.15 in. Example 4.8 showed that approximate maximum likelihood estimates for the gamma distribution characterizing January Ithaca precipitation, 1933–1982, are $\alpha = 3.76$ and $\beta = 0.52$ in., and that the cumulative probability corresponding to $x = 3.15$ in. in the context of this distribution is $F(3.15\,\text{in.}) = 0.873$. The SPI for the January 1987 precipitation at Ithaca is then the normal quantile transform (Equation 4.57) of the precipitation amount, $\text{SPI} = \Phi^{-1}[F(3.15\,\text{in.})] = \Phi^{-1}[0.873] = +1.14$. That is, the standard Gaussian variate having the same cumulative probability as does the 1987 January precipitation within its own climatological distribution is $z = +1.14$.

The SPI is routinely computed for timescales ranging from 1 month, as was just done for January 1987, through 2 years. For any timescale, the accumulated precipitation is characterized in terms of the corresponding cumulative probability within the distribution fitted to historical data over the same timescale and same portion of the annual cycle. So, for example, a 2-month SPI for January–February at a location of interest would involve fitting a gamma (or other suitable) distribution to the historical record of January plus February precipitation at that location. Similarly, annual SPI values are computed with respect to probability distributions for total annual precipitation. ◇

4.4.6. Beta Distributions

Some variables are restricted to segments of the real line that are bounded on two sides. Often such variables are restricted to the interval $0 \le x \le 1$. Examples of physically important variables subject to this restriction are cloud amount (observed as a fraction of the sky) and relative humidity. An important, more abstract, variable of this type is probability, where a parametric distribution can be useful in summarizing the frequency of use of forecasts, for example, of daily rainfall probability. The parametric distribution usually chosen to represent variations in these types of data is the *beta distribution*.

The PDF of the beta distribution is

$$f(x) = \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\right] x^{\alpha-1}\,(1-x)^{\beta-1}, \quad 0 \le x \le 1,\ \alpha,\beta > 0. \tag{4.58}$$

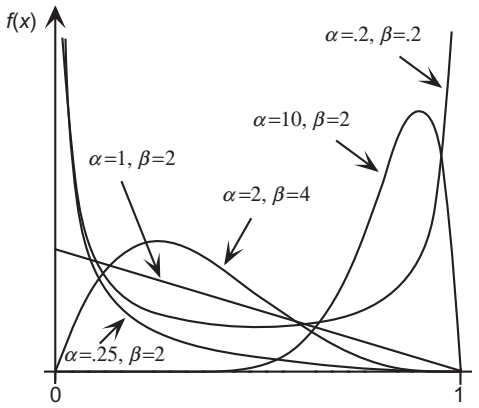

FIGURE 4.12 Five example PDFs for beta distributions. Mirror images of these distributions are obtained by reversing the parameters α and β.

This is a very flexible function, taking on many different shapes depending on the values of its two parameters, α and β. Figure 4.12 illustrates five of these. In general, for $\alpha \leq 1$ probability is concentrated near zero (e.g., $\alpha = 0.25$ and $\beta = 2$, or $\alpha = 1$ and $\beta = 2$, in Figure 4.12), and for $\beta \leq 1$ probability is concentrated near 1. If both parameters are less than one the distribution is U-shaped. For $\alpha > 1$ and $\beta > 1$ the distribution has a single mode (hump) between 0 and 1 (e.g., $\alpha = 2$ and $\beta = 4$, or $\alpha = 10$ and $\beta = 2$, in Figure 4.12), with more probability shifted to the right for $\alpha > \beta$, and more probability shifted to the left for $\alpha < \beta$. Beta distributions with $\alpha = \beta$ are symmetric. Reversing the values of α and β in Equation 4.58 results in a density function that is the mirror image (horizontally flipped) of the original.

Beta distribution parameters usually are fit using the method of moments. Using the expressions for the first two moments of the distribution,

$$\mu = \alpha/(\alpha+\beta) \tag{4.59a}$$

and

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{4.59b}$$

the moments estimators

$$\hat{\alpha} = \frac{\overline{x}^2(1-\overline{x})}{s^2} - \overline{x} \tag{4.60a}$$

and

$$\hat{\beta} = \frac{\hat{\alpha}(1-\overline{x})}{\overline{x}} \tag{4.60b}$$

are easily obtained. Alternatively, Mielke (1975) presents algorithms for maximum likelihood estimation.

The *uniform*, or *rectangular distribution*, is an important special case of the beta distribution, with $\alpha = \beta = 1$, and PDF $f(x) = 1$. The uniform distribution plays a central role in the computer generation of random numbers (see Section 4.7.1).

Use of the beta distribution is not limited only to variables having support on the unit interval [0, 1]. A variable, say y, constrained to any interval $[a, b]$ can be represented by a beta distribution after subjecting it to the transformation

$$x = \frac{y - a}{b - a}. \tag{4.61}$$

In this case parameter fitting is accomplished using

$$\overline{x} = \frac{\overline{y} - a}{b - a} \tag{4.62a}$$

and

$$s_x^2 = \frac{s_y^2}{(b - a)^2}, \tag{4.62b}$$

which are then substituted into Equation 4.60.

The integral of the beta probability density does not exist in closed form except for a few special cases, for example, the uniform distribution. Probabilities can be obtained through numerical methods (Abramowitz and Stegun 1984, Press et al. 1986), where the CDF for the beta distribution is known as the *incomplete beta function*, $I_x(\alpha,\beta) = \Pr\{0 \le X \le x\} = F(x)$. Tables of beta distribution probabilities are given in Epstein (1985) and Winkler (1972b).

4.4.7. Extreme-Value Distributions, I. Block-Maximum Statistics

The statistics of *extreme values* can be approached through description of the behavior of the largest of m values. Such data are extreme in the sense of being unusually large, and by definition are also rare. Often extreme-value statistics are of interest because the physical processes generating extreme events, and the societal impacts that occur because of them, are also large and unusual. A typical example of block-maximum extreme-value data is the collection of *annual maximum* daily precipitation values. In each of n years there is a wettest day of the $m = 365$ days, and the collection of these n wettest days is an extreme-value data set. Table 4.6 presents a small example annual maximum data set, for daily precipitation at Charleston, South Carolina. For each of the $n = 20$ years, the precipitation amount for the wettest of its $m = 365$ days is shown in the table.

A basic result from the theory of extreme-value statistics states (e.g., Coles 2001, Leadbetter et al. 1983) that the largest of m independent observations from a fixed distribution will follow a known

TABLE 4.6 Annual Maxima of Daily Precipitation Amounts (Inches) at Charleston, South Carolina, 1951–1970

1951	2.01	1956	3.86	1961	3.48	1966	4.58
1952	3.52	1957	3.31	1962	4.60	1967	6.23
1953	2.61	1958	4.20	1963	5.20	1968	2.67
1954	3.89	1959	4.48	1964	4.93	1969	5.24
1955	1.82	1960	4.51	1965	3.50	1970	3.00

distribution increasingly closely as m increases, regardless of the (single, fixed) distribution from which the observations have come. This result is called the *Extremal Types Theorem* and is the analog within the statistics of extremes of the Central Limit Theorem for the distribution of sums converging to the Gaussian distribution. The Extremal Types Theorem is valid even if the data within blocks are not independent, although in that case effectively the blocks are shorter (Coles 2001). The theory and approach are equally applicable to distributions of extreme minima (smallest of m observations) by analyzing the variable $-X$.

The distribution toward which the sampling distributions of largest-of-m values converges is called the *generalized extreme value*, or GEV, distribution, with PDF

$$f(x) = \frac{1}{\beta}\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{-1-\frac{1}{\kappa}} \exp\left\{-\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{-\frac{1}{\kappa}}\right\}, \quad 1 + \kappa(x-\zeta)/\beta > 0. \tag{4.63}$$

Here there are three parameters: a location (or shift) parameter ζ, a scale parameter β, and a shape parameter κ. Equation 4.63 can be integrated analytically, yielding the CDF

$$F(x) = \exp\left\{-\left[1 + \frac{\kappa(x-\zeta)}{\beta}\right]^{-\frac{1}{\kappa}}\right\}, \tag{4.64}$$

and this CDF can be inverted to yield an explicit formula for the quantile function,

$$F^{-1}(p) = \zeta + \frac{\beta}{\kappa}\{[-\ln(p)]^{-\kappa} - 1\}. \tag{4.65}$$

Particularly in the hydrological literature, Equations 4.63 through 4.65 are often written with the sign of the shape parameter κ reversed.

Because the moments of the GEV (see Table 4.5) involve the gamma function, estimating GEV parameters using the method of moments is no more convenient than alternative methods that yield more accurate results. The distribution usually is fit using either the method of maximum likelihood (see Section 4.6), or a method known as *L-moments* (Hosking 1990, Stedinger et al. 1993) that is used frequently in hydrological applications. L-moments fitting tends to be preferred for small data samples (Hosking 1990). Maximum likelihood methods can be adapted easily to include effects of *covariates* or additional influences. For example, the possibility that one or more of the distribution parameters may have a trend due to climate changes can be represented easily (Cooley 2009, Katz et al. 2002, Kharin and Zwiers 2005, Lee et al. 2014, Zhang et al. 2004). For moderate and large sample sizes the results of the two parameter estimation methods are usually similar. Using the data in Table 4.6, the maximum likelihood estimates for the GEV parameters are $\zeta = 3.50$, $\beta = 1.11$, and $\kappa = -0.29$, and the corresponding L-moment estimates are $\zeta = 3.49$, $\beta = 1.18$, and $\kappa = -0.32$.

Gumbel Distribution

Three special cases of the GEV are recognized, depending on the value of the shape parameter κ. The limit of Equation 4.63 as κ approaches zero yields the PDF

$$f(x) = \frac{1}{\beta} \exp\left\{-\exp\left[-\frac{(x-\zeta)}{\beta}\right] - \frac{(x-\zeta)}{\beta}\right\}, \tag{4.66}$$

known as the *Gumbel*, or *Fisher–Tippett Type I*, *distribution*. The Gumbel distribution is the limiting form of the GEV for extreme data drawn independently from distributions with well-behaved (i.e., exponential) tails, such as the Gaussian and the gamma. However, it is not unusual to find that the right tail of the Gumbel distribution may be too thin for this distribution to appropriately represent probabilities for daily rainfall extremes (e.g., Brooks and Carruthers 1953). The Gumbel distribution is so frequently used to represent the statistics of extremes that it is sometimes incorrectly called "the" extreme-value distribution. The Gumbel PDF is skewed to the right and exhibits its maximum at $x = \zeta$. Gumbel distribution probabilities can be obtained from the CDF

$$F(x) = \exp\left\{-\exp\left[-\frac{(x-\zeta)}{\beta}\right]\right\}. \tag{4.67}$$

Gumbel distribution parameters can be estimated through maximum likelihood or L-moments, as described earlier for the more general case of the GEV, but the simplest way to fit this distribution is to use the method of moments. The moments estimators for the two Gumbel distribution parameters are computed using the sample mean and standard deviation. The estimation equations are

$$\hat{\beta} = \frac{s\sqrt{6}}{\pi} \tag{4.68a}$$

and

$$\hat{\zeta} = \bar{x} - \gamma\hat{\beta}, \tag{4.68b}$$

where $\gamma = 0.57721\ldots$ is *Euler's constant*.

Fréchet Distribution

For $\kappa > 0$ the Equation 4.63 is called the *Fréchet*, or *Fisher–Tippett Type II distribution*.

These distributions exhibit what are called "heavy" tails, meaning that the PDF decreases rather slowly for large values of x. One consequence of heavy tails is that some of the moments of Frechet distributions are not finite. For example, the integral defining the variance (Equation 4.22) is infinite for $\kappa > 1/2$, and even the mean (Equation 4.21 with $g(x) = x$) is not finite for $\kappa > 1$. The notable practical consequence of heavy tails is that quantiles associated with large cumulative probabilities (i.e., Equation 4.65 with $p \approx 1$) will be quite large. It is often found that annual maxima for precipitation and streamflow data exhibit heavy tails (e.g., Morrison and Smith 2002, Papalexiou and Koutsoyiannis 2013, Serinaldi and Kilsby 2014).

Weibull Distribution

The third special case of the GEV distribution occurs for $\kappa < 0$ and is known as the *Weibull*, or *Fisher–Tippett Type III distribution*. In addition to use in extreme-value statistics, Weibull distributions are often chosen to represent wind data (e.g., Conradsen et al. 1984, He et al. 2010, Justus et al., 1978). Typically Weibull distributions are written with the shift parameter $\zeta = 0$, and a parameter transformation, yielding the PDF

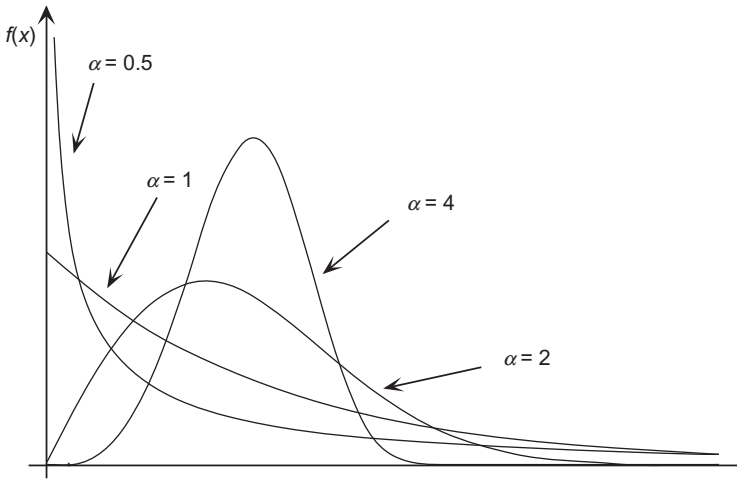

FIGURE 4.13 Weibull distribution PDFs for four values of the shape parameter, α.

$$f(x) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right], \quad x, \alpha, \beta > 0. \tag{4.69}$$

As is the case for the gamma distribution, the two parameters α and β are called the shape and scale parameters, respectively. The form of the Weibull distribution also is controlled similarly by the two parameters. The response of the shape of the distribution to different values of α is shown in Figure 4.13. In common with the gamma distribution, $\alpha \leq 1$ produces reverse "J" shapes and strong positive skewness, and for $\alpha = 1$ the Weibull distribution also reduces to the exponential distribution (Equation 4.52) as a special case. Also in common with the gamma distribution, the scale parameter acts similarly to either stretch or compress the basic shape along the x-axis, for a given value of α. For $\alpha \approx 3.6$ the Weibull is very similar to the Gaussian distribution. However, for shape parameters larger than this the Weibull density exhibits negative skewness, which is visible in Figure 4.13 for $\alpha = 4$.

The PDF for the Weibull distribution is analytically integrable, resulting in the CDF

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]. \tag{4.70}$$

This equation can easily be solved for x to yield the quantile function. As is the case for the GEV more generally, the moments of the Weibull distribution involve the gamma function (see Table 4.5), so there is no computational advantage to parameter fitting by the method of moments. Usually Weibull distributions are fit using either maximum likelihood (see Section 4.6) or L-moments (Stedinger et al. 1993).

Average Return Periods

The result of an extreme-value analysis is often simply a summary of quantiles corresponding to large cumulative probabilities, for example, the event with an annual probability of 0.01 of being exceeded. Unless n is rather large, direct empirical estimation of these extreme quantiles will not be possible (cf. Equation 3.19), and a well-fitting extreme-value distribution provides a reasonable and objective way to

extrapolate to probabilities that may be substantially larger than $1 - 1/n$. Often these extreme probabilities are expressed as average *return periods*,

$$R(x) = \frac{1}{\omega\,[1 - F(x)]}. \tag{4.71}$$

The return period $R(x)$ associated with a quantile x typically is interpreted to be the average time between occurrences of events of that magnitude or greater. The return period is a function of the CDF evaluated at x and the average sampling frequency ω. For annual maximum data $\omega = 1$ per year, in which case the event x corresponding to a cumulative probability $F(x) = 0.99$ will have probability $1 - F(x)$ of being exceeded in any given year. This value of x would be associated with a return period of 100 years and would be called the 100-year event. Extreme-event risks expressed in terms of return periods appear to be less well understood by the general public than are annual event probabilities (Grounds et al., 2018). In addition, the computation of return periods in Equation 4.71 implicitly assumes climate stationarity, but as the climate changes the (e.g.) 100-year event may become larger or smaller over the next century. It may thus be better to express extreme-value risk as the annual occurrence probability $F(x)$, changes in which might be modeled using time trends for the distribution parameters (e.g., Chen and Chu, 2014; Katz et al. 2002; Kharin and Zwiers, 2005).

Example 4.11. Return Periods and Cumulative Probability

As noted earlier, a maximum likelihood fit of the GEV distribution to the annual maximum precipitation data in Table 4.6 yielded the parameter estimates $\zeta = 3.50$, $\beta = 1.11$, and $\kappa = -0.29$. Using Equation 4.65 with cumulative probability $p = 0.5$ yields a median of 3.89 in. This is the daily precipitation amount that has a 50% chance of being exceeded in a given year. This amount will therefore be exceeded on average in half of the years in a hypothetical long climatological record, and so the average time separating daily precipitation events of this magnitude or greater is 2 years (Equation 4.71).

Because $n = 20$ years for these data, the median can be well estimated directly as the sample median. But consider estimating the 100-year 1-day precipitation event from these data. According to Equation 4.71 this corresponds to the cumulative probability $F(x) = 0.99$, whereas the empirical cumulative probability corresponding to the most extreme precipitation amount in Table 4.6 might be estimated as $p \approx 0.967$, using the Tukey plotting position (see Table 3.2). However, using the GEV quantile function (Equation 4.65) together with Equation 4.71, a reasonable estimate for the 100-year maximum daily amount is calculated to be 6.32 in. The corresponding 2- and 100-year precipitation amounts derived from the L-moment parameter estimates, $\zeta = 3.49$, $\beta = 1.18$, and $\kappa = -0.32$, are 3.90 and 6.33 in., respectively.

It is worth emphasizing that the T-year event is in no way guaranteed to occur within a particular period of T years, and indeed the probability distribution for the waiting time until the next occurrence of an extreme event will be quite broad (e.g., Wigley 2009). The probability that the T-year event occurs in any given year is $1/T$, for example, $1/T = 0.01$ for the $T = 100$-year event. In any particular year, the occurrence of the T-year event is a Bernoulli trial, with $p = 1/T$. Therefore the geometric distribution (Equation 4.5) can be used to calculate probabilities of waiting particular numbers of years for the event. Another interpretation of the return period is as the mean of the geometric distribution for the waiting time. The probability of the 100-year event occurring in an arbitrarily chosen century can be calculated as $\Pr\{X \le 100\} = 0.634$ using Equation 4.5. That is, there is more than a 1/3 chance that the 100-year event will not occur in any particular 100 years. Similarly, the probability of the 100-year event not occurring in 200 years is approximately 0.134. ◇

One important motivation for studying and modeling the statistics of extremes is to estimate annual probabilities of rare and potentially damaging events, such as extremely large daily precipitation amounts that might cause flooding, or extremely large wind speeds that might cause damage to structures. In applications like these, the assumptions of classical extreme-value theory, namely, that the underlying events are independent and come from the same distribution, and that the number of individual (usually daily) values m is sufficient for convergence to the GEV, may not be met. Most problematic for the application of extreme-value theory is that the underlying data often will not be drawn from the same distribution, for example, because of an annual cycle in the statistics of the m (=365, usually) values, and/or because the largest of the m values are generated by different processes in different blocks (years). For example, some of the largest daily precipitation values may occur because of hurricane landfalls, some may occur because of large and slowly moving thunderstorm complexes, and some may occur as a consequence of near-stationary frontal boundaries. The statistics of (i.e., the underlying PDFs corresponding to) the different physical processes may be different (e.g., Walshaw 2000).

These considerations do not invalidate the GEV (Equation 4.63) as a candidate distribution to describe the statistics of extremes, and empirically this distribution often is found to be an excellent choice even when the assumptions of extreme-value theory are not met. However, in the many practical settings where the classical assumptions are not valid the GEV is not guaranteed to be the most appropriate distribution to represent a set of extreme-value data. The appropriateness of the GEV should be evaluated along with other candidate distributions for particular data sets (Madsen et al. 1997, Wilks 1993), possibly using approaches presented in Section 4.6 or 5.2.6.

4.4.8. Extreme-Value Distributions, II. Peaks-Over-Threshold Statistics

Approaching the statistics of extremes using the block-maximum approach of Section 4.4.7 can be wasteful of data, when there are values that are not largest in their block (e.g., year of occurrence) but may be larger than the maxima in other blocks. An alternative approach to assembling a set of extreme-value data is to choose the largest n values regardless of their year of occurrence. The result is called *partial-duration* data in hydrology. This approach is known more generally as *peaks-over-threshold*, or POT, since any values larger than a (large) minimum level are chosen, and we are not restricted to choosing the same number of extreme values as there may be years in the climatological record. Because the underlying data may exhibit substantial serial correlation, some care is required to ensure that selected partial-duration data represent distinct events.

When the data underlying an extreme-value analysis have been abstracted using POT sampling, there is some theoretical support for characterizing them using the *generalized Pareto distribution*, with PDF

$$f(x) = \frac{1}{\sigma^*}\left[1 + \frac{\kappa(x-u)}{\sigma^*}\right]^{-\frac{1}{\kappa}-1} \tag{4.72}$$

and CDF

$$F(x) = 1 - \left[1 + \frac{\kappa(x-u)}{\sigma^*}\right]^{-1/\kappa}. \tag{4.73}$$

This distribution arises as an approximation to the distribution of POT data taken from a GEV distribution, with the numbers of these peaks in a given time period following a Poisson distribution (Coles 2001, Hosking and Wallis 1987, Katz et al. 2002). Here u is the threshold for the POT sampling,

which should be relatively high. The shape parameter κ in Equations 4.72 and 4.73 has the same value as that for the related GEV distribution. Also in common with the GEV, generalized Pareto distributions with $\kappa > 0$ exhibit heavy tails. The scale parameter σ^* in Equations 4.72 and 4.73 is related to the parameters of the corresponding GEV distribution and to the sampling threshold u according to

$$\sigma^* = \beta + \kappa(u - \zeta). \tag{4.74}$$

Coles (2001) suggests that u can be optimized by fitting the distribution using a range of thresholds, and choosing a value slightly above the point at which the fitted shape parameters κ are fairly constant, and the scale parameters σ^* vary linearly according to Equation 4.74. An alternative empirical choice that has been suggested in the context of daily data is to choose u such that $1.65n$ extreme values are chosen, given n years of data (Madsen et al. 1997, Stedinger et al. 1993), yielding $\omega = 1.65\,\text{year}^{-1}$ in Equation 4.71.

The foregoing theory assumes statistical independence of the POT data, whereas often serial correlation in the underlying data series will give rise to clusters of consecutive values larger than u. Each of these clusters might reasonably be interpreted as a single event. Coles (2001) describes a process of *declustering*, whereby a subjective empirical rule defines clusters of exceedances and only the largest value within each cluster is incorporated into the POT data set. Nonstationarity (e.g., trends due to climate change) can be dealt with in POT settings with the same approach taken for block-maximum data, namely, by expressing one or more of the distribution parameters as a function of time (e.g., Katz 2013).

4.4.9. Mixture Distributions

The parametric distributions presented so far in this chapter may be inadequate for data that arise from more than one generating process or physical mechanism. An example is the Guayaquil temperature data in Table A.3, for which histograms are shown in Figure 3.6. These data are clearly bimodal; with the smaller, warmer hump in the distribution associated with El Niño years, and the larger, cooler hump consisting mainly of the non-El Niño years. Although the Central Limit Theorem suggests that the Gaussian distribution should be a good model for monthly averaged temperatures, the clear differences in the Guayaquil June temperature climate associated with El Niño make the Gaussian a poor choice to represent these data overall. However, separate Gaussian distributions for El Niño years and non-El Niño years might provide a good probability model for these data.

Cases like this are natural candidates for representation with *mixture distributions*, which are weighted averages of two or more PDFs. Any number of PDFs can be combined to form a mixture distribution (Everitt and Hand 1981, McLachlan and Peel 2000, Titterington et al. 1985), but by far the most commonly used mixture distributions are weighted averages of two component PDFs,

$$f(x) = w f_1(x) + (1 - w) f_2(x). \tag{4.75}$$

The component PDFs $f_1(x)$ and $f_2(x)$ can be any distributions, and usually although not necessarily they are of the same parametric form. The weighting parameter w, $0 < w < 1$, determines the contribution of each component density to the mixture PDF and can be interpreted as the probability that a realization of the random variable X will have come from $f_1(x)$.

Of course the properties of a mixture distribution depend on the properties of the component distributions and on the weighting parameter. The mean is simply the weighted average of the two component means,

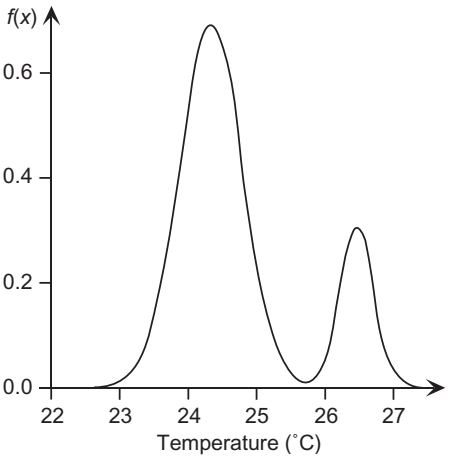

FIGURE 4.14 Probability density function for the mixture (Equation 4.75) of two Gaussian distributions fit to the June Guayaquil temperature data (Table A.3). The result is very similar to the kernel density estimate derived from the same data, Figure 3.8b.

$$\mu = w\mu_1 + (1-w)\mu_2. \tag{4.76}$$

On the other hand, the variance

$$\begin{aligned}\sigma^2 &= \left[w\sigma_1^2 + (1-w)\sigma_2^2\right] + \left[w(\mu_1-\mu)^2 + (1-w)(\mu_2-\mu)^2\right] \\ &= w\sigma_1^2 + (1-w)\sigma_2^2 + w(1-w)(\mu_1-\mu_2)^2,\end{aligned} \tag{4.77}$$

has contributions from the weighted variances of the two distributions (first square-bracketed terms on the first line), plus additional dispersion deriving from the difference of the two means (second square-bracketed terms). Mixture distributions are clearly capable of representing bimodality (or, when the mixture is composed of three or more component distributions, multimodality), but mixture distributions can also be unimodal if the differences between component means are small enough relative to the component standard deviations or variances.

Usually mixture distributions are fit using maximum likelihood, using the EM algorithm (see Section 4.6.3). Figure 4.14 shows the PDF for a maximum likelihood fit of a mixture of two Gaussian distributions to the June Guayaquil temperature data in Table A.3, with parameters $\mu_1 = 24.34°C$, $\sigma_1 = 0.46°C$, $\mu_2 = 26.48°C$, $\sigma_2 = 0.26°C$, and $w = 0.80$ (see Example 4.14). Here μ_1 and σ_1 are the parameters of the first (cooler and more probable) Gaussian distribution, $f_1(x)$, and μ_2 and σ_2 are the parameters of the second (warmer and less probable) Gaussian distribution, $f_2(x)$. The mixture PDF in Figure 4.14 results as a simple (weighted) addition of the two component Gaussian distributions, in a way that is similar to the construction of the kernel density estimates for the same data in Figure 3.8, as a sum of scaled kernels that are themselves PDFs. Indeed, the Gaussian mixture in Figure 4.14 resembles the kernel density estimate derived from the same data in Figure 3.8b. The means of the two component Gaussian distributions are well separated relative to the dispersion characterized by the two standard deviations, resulting in the mixture distribution being strongly bimodal.

Gaussian distributions are the most common choice for components of mixture distributions, but mixtures of exponential distributions (Equation 4.52) are also important and frequently used. In particular, the mixture distribution composed of two exponential distributions is called the *mixed exponential distribution* (Smith and Schreiber 1974), with PDF

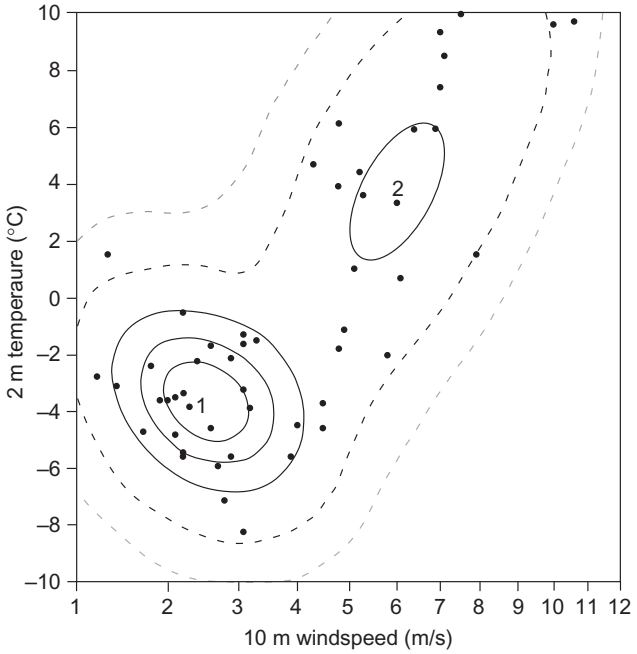

FIGURE 4.15 Contour plot of the PDF of a bivariate Gaussian mixture distribution, fit to an ensemble of 51 forecasts for 2 m temperature and 10 m wind speed, made at 180 h lead time. The wind speeds have first been square root transformed to make their univariate distribution more Gaussian. Dots indicate individual forecasts made by the 51 ensemble members. The two constituent bivariate Gaussian densities $f_1(x)$ and $f_2(x)$ are centered at "1" and "2," respectively, and the smooth lines indicate level curves of their mixture $f(x)$, formed with $w = 0.57$. Solid contour interval is 0.05, and the heavy and light dashed lines are 0.01 and 0.001, respectively. Adapted from Wilks (2002b).

$$f(x) = \frac{w}{\beta_1}\exp\left(-\frac{x}{\beta_1}\right) + \frac{1-w}{\beta_2}\exp\left(-\frac{x}{\beta_2}\right). \tag{4.78}$$

The mixed exponential distribution has been found to be well suited for representing nonzero daily precipitation data (Woolhiser and Roldan 1982, Foufoula-Georgiou and Lettenmaier 1987, Wilks 1999a) and is especially useful for simulating (see Section 4.7) spatially correlated daily precipitation amounts (Wilks 1998).

Mixture distributions are not limited to combinations of univariate continuous PDFs. Equation 4.75 can as easily be used to form mixtures of discrete probability distribution functions or mixtures of multivariate joint distributions. For example, Figure 4.15 shows the mixture of two bivariate Gaussian distributions (Equation 4.31) fit to a 51-member ensemble forecast (see Section 8.2) for temperature and wind speed. The distribution was fit using the maximum likelihood algorithm for multivariate Gaussian mixtures given in Smyth et al. (1999) and Hannachi and O'Neill (2001). Although multivariate mixture distributions are quite flexible in accommodating unusual-looking data, this flexibility comes at the cost of needing to estimate a large number of parameters, so use of relatively elaborate probability models of this kind may be limited by the available sample size. The mixture distribution in Figure 4.15 requires 11 parameters to characterize it: two means, two variances, and one correlation for each of the two component bivariate distributions, plus the weight parameter w.

Other Combination Distributions

Combinations of component distributions of different parametric types have been used especially to represent hydrologic variables, in order to better model their extreme tail behavior. Although not a mixture distribution in the sense of Equation 4.75, Furrer and Katz (2008) combined gamma and generalized Pareto distributions to represent daily precipitation amounts, by stitching together a gamma distribution for most of the data with a generalized Pareto distribution for the extreme amounts. The generalized Pareto scale parameter was adjusted in order to achieve a continuous (although not smooth) transition between the two. Carreau and Bengio (2009) take a similar approach when augmenting Gaussian distributions with generalized Pareto tails, and Li et al. (2012) analogously combine exponential distributions with generalized Pareto tails.

Vrac and Naveau (2007) employed a "dynamic" mixture of gamma and generalized Pareto distributions, in which the weight w in Equation 4.75 depends on the amount x, so that the mixture density is defined as

$$f(x) = \frac{[1 - w(x)]f_\Gamma(x) + w(x)f_{gP}(x)}{c}, \tag{4.79}$$

where $f_\Gamma(x)$ denotes a gamma density (Equation 4.45), $f_{gP}(x)$ denotes a generalized Pareto density (Equation 4.72) with threshold $u = 0$, and c is a normalizing constant which has been defined explicitly by Li et al. (2012). The weight function takes the form

$$w(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x - m}{\tau}\right), \tag{4.80}$$

where m and τ are fitted parameters. Equation 4.80 obviates the need to choose a threshold u, progressively downweights the light-tailed gamma distribution in favor of the generalized Pareto tail distribution as x increases, and ensures a smooth and continuous transition between the two.

4.5. QUALITATIVE ASSESSMENTS OF THE GOODNESS OF FIT

Having fit a parametric distribution to a batch of data, it is worthwhile to verify that the theoretical probability model provides an adequate description. Fitting an inappropriate distribution can lead to erroneous conclusions being drawn. Quantitative methods for evaluating the closeness of fitted distributions to underlying data rely on ideas from formal hypothesis testing, and a few such methods will be presented in Section 5.2.5. This section describes some qualitative, graphical methods useful for subjectively discerning the goodness of fit. These methods are instructive even if a formal goodness of fit test of fit is to be computed also. A formal test may indicate an inadequate fit, but it may not inform the analyst as to the specific nature of the discrepancies. Graphical comparison of the data and the fitted distribution allow diagnosis of where and how the parametric representation may be inadequate.

4.5.1. Superposition of a Fitted Parametric Distribution and Data Histogram

Superposition of the fitted distribution and the histogram is probably the simplest and most intuitive means of comparing a fitted parametric distribution to the underlying data. Gross departures of the parametric model from the data can readily be seen in this way. If the data are sufficiently numerous, irregularities in the histogram due to sampling variations will not be too distracting.

For discrete data, the probability distribution function is already very much like the histogram. Both the histogram and the probability distribution function assign probability to a discrete set of outcomes. Comparing the two requires only that the same discrete data values, or ranges of the data values, are plotted, and that the histogram and distribution function are scaled comparably. This second condition is met by plotting the histogram in terms of relative, rather than absolute, frequency on the vertical axis. Figure 4.2 is an example, showing the superposition of a Poisson probability distribution function on the histogram of observed annual numbers U.S. hurricane landfalls.

The procedure for superimposing a continuous PDF on a histogram is entirely analogous. The fundamental constraint is that the integral of any PDF, over the full range of the random variable, must be one. That is, Equation 4.18 is satisfied by all PDFs. One approach to matching the histogram and the density function is to rescale the density function. The correct scaling factor is obtained by computing the area occupied collectively by all the bars in the histogram plot. Denoting this area as A, it is easy to see that multiplying the fitted density function $f(x)$ by A produces a curve whose area is also A because, as a constant, A can be taken out of the integral: $\int_x A \cdot f(x)\, dx = A \cdot \int_x f(x)\, dx = A \cdot 1 = A$. Note that it is also possible to rescale the histogram heights so that the total area contained in the bars is 1. This latter approach is more traditional in statistics, since the histogram is regarded as an estimate of the density function.

Example 4.12. Superposition of PDFs onto a Histogram

Figure 4.16 illustrates the procedure of superimposing fitted distributions and a histogram, for the 1933–1982 January precipitation totals at Ithaca from Table A.2. Here $n = 50$ years of data, and the bin width for the histogram (consistent with Equation 3.13) is 0.5 in., so the area occupied by the histogram rectangles is $A = (50)(0.5) = 25$. Superimposed on this histogram are PDFs for the gamma distribution fit

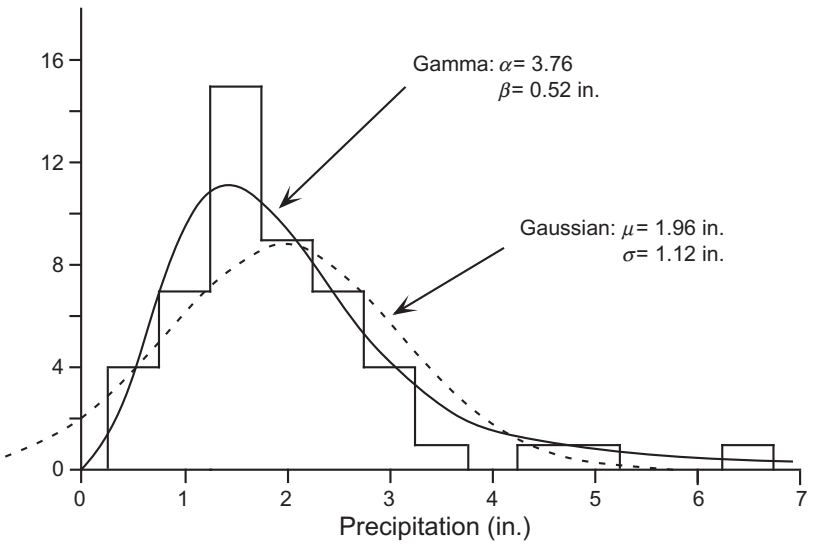

FIGURE 4.16 Histogram of the 1933–1982 Ithaca January precipitation data from Table A.2, with the fitted gamma *(solid)* and Gaussian *(dashed)* PDFs. Each of the two density functions has been multiplied by $A = 25$, since the bin width is 0.5 in. and there are 50 observations. Apparently the gamma distribution provides a reasonable representation of the data. The Gaussian distribution underrepresents the right tail and implies nonzero probability for negative precipitation.

using Equation 4.48 or 4.50a *(solid curve)*, and the Gaussian distribution fit by matching the sample and distribution moments *(dashed curve)*. In both cases the PDFs (Equations 4.45 and 4.24, respectively) have been multiplied by 25 so that their areas are equal to that of the histogram. It is clear that the symmetrical Gaussian distribution is a poor choice for representing these positively skewed precipitation data, since too little probability is assigned to the largest precipitation amounts and nonnegligible probability is assigned to impossible negative precipitation amounts. The gamma distribution represents these data much more closely and provides a quite plausible summary of the year-to-year variations in the data. The fit appears to be worst for the 0.75–1.25 in. and 1.25–1.75 in. bins, although this easily could have resulted from sampling variations. This same data set will also be used in Section 5.2.5 to test formally the fit of these two distributions. ◇

4.5.2. Quantile–Quantile (Q–Q) Plots

Quantile–quantile (*Q–Q*) *plots* compare empirical (data) and fitted CDFs in terms of the dimensional values of the variable (the empirical quantiles). The link between observations of the random variable X and the fitted distribution is made through the quantile function, or inverse of the CDF (Equation 4.20), evaluated at estimated levels of cumulative probability.

The Q–Q plot is a scatterplot. Each coordinate pair defining the location of a point consists of a data value and the corresponding estimate for that data value derived from the quantile function of the fitted distribution. Adopting the Tukey plotting position formula (see Table 3.2) as the estimator for empirical cumulative probability (although others could reasonably be used), each point in a Q–Q plot would have the Cartesian coordinates $(F^{-1}[(i-1/3)/(n+1/3)], x_{(i)})$. Thus the ith point on the Q–Q plot is defined by the ith smallest data value, $x_{(i)}$, and the value of the random variable corresponding to the sample cumulative probability $p = (i-1/3)/(n+1/3)$ in the fitted distribution. A Q–Q plot for a fitted distribution representing the data perfectly would have all points falling on the 1:1 diagonal line.

Figure 4.17 shows Q–Q plots comparing the fits of gamma and Gaussian distributions to the 1933–1982 Ithaca January precipitation data in Table A.2 (the parameter estimates are shown in Figure 4.16). Figure 4.17 indicates that the fitted gamma distribution corresponds well to the data through most of its range, since the quantile function evaluated at the estimated empirical cumulative probabilities is quite close to the observed data values, yielding points very close to the diagonal 1:1 line. The fitted distribution seems to underestimate the largest few points, suggesting that the tail of the fitted gamma distribution may be too thin.

On the other hand, Figure 4.17 shows the Gaussian fit to these data is clearly inferior. Most prominently, the left tail of the fitted Gaussian distribution is too heavy, so that the smallest theoretical quantiles are too small, and in fact the smallest two are actually negative. Through the bulk of the distribution the Gaussian quantiles are further from the 1:1 line than the gamma quantiles, indicating a less accurate fit, and on the right tail the Gaussian distribution underestimates the largest quantiles even more than does the gamma distribution.

It is possible also to compare fitted and empirical distributions by reversing the logic of the Q–Q plot, and producing a scatterplot of the empirical cumulative probability (estimated using a plotting position, Table 3.2) as a function of the fitted CDF, $F(x)$, evaluated at the corresponding data value. Plots of this kind are called *probability–probability*, or *P–P plots*. P–P plots seem to be used less frequently than Q–Q plots, perhaps because comparisons of dimensional data values can be more intuitive than comparisons of cumulative probabilities. Also, because P–P plots converge to the (0,0) and (1,1) points at the extremes, regardless of the closeness of correspondence, they are less sensitive to differences in the tails of a

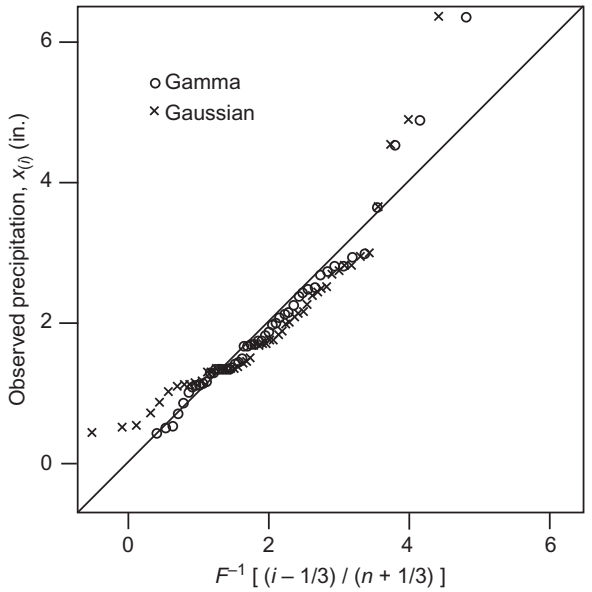

FIGURE 4.17 Quantile–quantile plots for gamma (○) and Gaussian (×) fits to the 1933–1982 Ithaca January precipitation data in Table A.2. Observed precipitation amounts are on the vertical, and amounts inferred from the fitted distributions using the Tukey plotting position are on the horizontal. Diagonal line indicates 1:1 correspondence.

distribution, which are often of most interest. Both Q–Q and P–P plots belong to a broader class of plots known as *probability plots*.

4.6. PARAMETER FITTING USING MAXIMUM LIKELIHOOD

4.6.1. The Likelihood Function

For many distributions, fitting parameters using the simple method of moments produces inferior results that can lead to misleading inferences and extrapolations. The *method of maximum likelihood* is a versatile and important alternative. As the name suggests, the method seeks to find values of the distribution parameters that maximize the *likelihood function*. The procedure follows from the notion that the likelihood is a measure of the degree to which the data support particular values of the parameter(s) (e.g., Lindgren 1976). As explained more fully in Chapter 6, a Bayesian interpretation of the procedure (except for small sample sizes) would be that the maximum likelihood estimators (MLEs) are the most probable values for the parameters, given the observed data.

Notationally, the likelihood function for a single observation, x, looks identical to the probability density (or, for discrete variables, the probability distribution) function, and the difference between the two can be confusing initially. The distinction is that the PDF is a function of the data for fixed values of the parameters, whereas the likelihood function is a function of the unknown parameters for fixed values of the (already observed) data. Just as the joint PDF of n independent variables is the product of the n individual PDFs, the likelihood function for the parameters of a distribution given a sample of n independent data values is the product of the n individual likelihood functions. For example,

the likelihood function for the Gaussian parameters μ and σ, given a sample of n observations, x_i, $i = 1, \ldots, n$, is

$$\Lambda(\mu, \sigma) = \sigma^{-n}\left(\sqrt{2\pi}\right)^{-n} \prod_{i=1}^{n} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]. \tag{4.81}$$

Here the $\prod$ indicates multiplication of terms of the form indicated to its right. Actually, the Gaussian likelihood can be any function proportional to Equation 4.81, so the constant factor involving the square root of 2π could have been omitted because it does not depend on either of the two parameters. It has been included here to emphasize the relationship between Equations 4.81 and 4.24. The right-hand side of Equation 4.81 looks exactly the same as the joint PDF for n independent Gaussian variates from the same distribution, except that the parameters μ and σ are the variables, and the x_i denote fixed constants. Geometrically, Equation 4.81 describes a surface above the μ–σ plane that takes on a maximum value above a specific pair of parameter values, which depend on the particular data set given by the x_i values.

Usually it is more convenient to work with the logarithm of the likelihood function, known as the *log-likelihood*. Since the logarithm is a strictly increasing function, the same parameter values will maximize both the likelihood and log-likelihood functions. The log-likelihood function for the Gaussian parameters, corresponding to Equation 4.81 is

$$L(\mu, \sigma) = \ln[\Lambda(\mu, \sigma)] = -n \ln(\sigma) - n \ln\left(\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2, \tag{4.82}$$

where, again, the term involving 2π is not strictly necessary for locating the maximum of the function because it does not depend on the parameters μ or σ.

Conceptually, at least, maximizing the log-likelihood is a straightforward exercise in calculus. For the Gaussian distribution the exercise really is straightforward, since the maximization can be done analytically. Taking derivatives of Equation 4.82 with respect to the parameters yields

$$\frac{\partial L(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n} x_i - n\mu\right] \tag{4.83a}$$

and

$$\frac{\partial L(\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{4.83b}$$

Setting each of these derivatives equal to zero and solving yields, respectively,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.84a}$$

and

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}. \tag{4.84b}$$

These are the MLEs for the Gaussian distribution, which are readily recognized as being very similar to the moments estimators. The only difference is the divisor in Equation 4.84b, which is n rather

than $n - 1$. The divisor $n - 1$ is often adopted when computing the sample standard deviation, because that choice yields an unbiased estimate of the population value. This difference points out the fact that the MLEs for a particular distribution may not be unbiased. In this case the estimated standard deviation (Equation 4.84b) will tend to be too small, on average, because the x_i are on average closer to the sample mean computed from them in Equation 4.84a than to the true mean, although these differences are small for large n.

4.6.2. The Newton–Raphson Method

The MLEs for the Gaussian distribution are somewhat unusual, in that they can be computed analytically. It is more usual for approximations to the MLEs to be calculated iteratively. One common approach is to think of the maximization of the log-likelihood as a nonlinear rootfinding problem to be solved using the multidimensional generalization of the *Newton–Raphson method* (e.g., Press et al. 1986). This approach follows from the truncated Taylor expansion of the derivative of the log-likelihood function

$$L'(\boldsymbol{\theta}^*) \approx L'(\boldsymbol{\theta}) + (\boldsymbol{\theta}^* - \boldsymbol{\theta})\, L''(\boldsymbol{\theta}), \tag{4.85}$$

where $\boldsymbol{\theta}$ denotes a generic vector of distribution parameters and $\boldsymbol{\theta}^*$ are the true values to be approximated. Since it is the *derivative* of the log-likelihood function, $L'(\boldsymbol{\theta}^*)$, whose roots are to be found, Equation 4.85 requires computation of the second derivatives of the log-likelihood, $L''(\boldsymbol{\theta})$. Setting Equation 4.85 equal to zero (to find a maximum in the log-likelihood, L) and rearranging yields the expression describing the algorithm for the iterative procedure,

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \frac{L'(\boldsymbol{\theta})}{L''(\boldsymbol{\theta})}. \tag{4.86}$$

Beginning with an initial guess, $\boldsymbol{\theta}$, an updated set of estimates $\boldsymbol{\theta}^*$ are computed by subtracting the ratio of the first to second derivatives, which are in turn used as the guesses for the next iteration.

Example 4.13. Algorithm for Maximum Likelihood Estimation of Gamma Distribution Parameters In practice, use of Equation 4.86 is somewhat complicated by the fact that usually more than one parameter must be estimated simultaneously, so that $L'(\boldsymbol{\theta})$ is a vector of first derivatives, and $L''(\boldsymbol{\theta})$ is a matrix of second derivatives. To illustrate, consider the gamma distribution (Equation 4.45). For this distribution, Equation 4.86 becomes

$$\begin{aligned}\begin{bmatrix}\alpha^* \\ \beta^*\end{bmatrix} &= \begin{bmatrix}\alpha \\ \beta\end{bmatrix} - \begin{bmatrix}\partial^2 L/\partial\alpha^2 & \partial^2 L/\partial\alpha\partial\beta \\ \partial^2 L/\partial\beta\partial\alpha & \partial^2 L/\partial\beta^2\end{bmatrix}^{-1} \begin{bmatrix}\partial L/\partial\alpha \\ \partial L/\partial\beta\end{bmatrix} \\ &= \begin{bmatrix}\alpha \\ \beta\end{bmatrix} - \begin{bmatrix}-n\Gamma''(\alpha) & -n/\beta \\ -n/\beta & \dfrac{n\alpha}{\beta^2} - \dfrac{2\sum x}{\beta^3}\end{bmatrix}^{-1} \begin{bmatrix}\sum \ln(x) - n\ \ln(\beta) - n\Gamma'(\alpha) \\ \sum x/\beta^2 - n\alpha/\beta\end{bmatrix},\end{aligned} \tag{4.87}$$

where $\Gamma'(\alpha)$ and $\Gamma''(\alpha)$ are the first and second derivatives of the gamma function (Equation 4.7), which must be evaluated or approximated numerically (e.g., Abramowitz and Stegun 1984). The matrix-algebra notation in this equation is explained in Chapter 11. Equation 4.87 would be implemented by starting with initial guesses for the parameters α and β, perhaps using the moments estimators (Equations 4.46). Updated values, α^* and β^*, would then result from a first application of Equation 4.87. The updated values would then be substituted into the right-hand side of Equation 4.87, and the process

repeated until convergence of the algorithm. Convergence might be diagnosed by the parameter estimates changing sufficiently little, perhaps by a small fraction of a percent, between iterations. Note that in practice the Newton–Raphson algorithm may overshoot the likelihood maximum on a given iteration, which could result in a decline from one iteration to the next in the current approximation to the log-likelihood. Often the Newton–Raphson algorithm is programmed in a way that checks for such likelihood decreases and tries smaller changes in the estimated parameters (although in the same direction specified by, in this case, Equation 4.87). ◇

4.6.3. The EM Algorithm

Maximum likelihood calculations using the Newton–Raphson method are generally fast and effective in applications where estimation of relatively few parameters is required. However, for problems involving more than perhaps three parameters, the computations required can expand dramatically. Even worse, the iterations can be quite unstable (sometimes producing "wild" updated parameters $\boldsymbol{\theta}^*$ well away from the maximum likelihood values being sought) unless the initial guesses are so close to the correct values that the estimation procedure itself is almost unnecessary.

The *EM*, or *Expectation–Maximization algorithm* (McLachlan and Krishnan 1997), is an alternative to Newton–Raphson that does not suffer these problems. It is actually somewhat imprecise to call the EM algorithm an "algorithm," in the sense that there is not an explicit specification (like Equation 4.86 for the Newton–Raphson method) of the steps required to implement it in a general way. Rather, it is more of a conceptual approach that needs to be tailored to particular problems.

The EM algorithm is formulated in the context of parameter estimation given "incomplete" data. Accordingly, on one level, it is especially well suited to situations where some data may be missing (censored), or unobserved (truncated) above or below known thresholds, or recorded imprecisely because of coarse binning. Such situations are handled easily by the EM algorithm when the estimation problem would be easy (e.g., reducing to an analytic solution such as Equation 4.84) if the data were "complete." More generally, an ordinary (i.e., not intrinsically "incomplete") estimation problem can be approached with the EM algorithm if the existence of some additional unknown (and possibly hypothetical or unknowable) data would allow formulation of a straightforward (e.g., analytical) maximum likelihood estimation procedure. Like the Newton–Raphson method, the EM algorithm requires iterated calculations, and therefore an initial guess for the parameters to be estimated. When the EM algorithm can be formulated for a maximum likelihood estimation problem, some of the difficulties experienced by the Newton–Raphson approach do not occur, and in particular the updated log-likelihood will not decrease from iteration to iteration, regardless of how many parameters are being estimated simultaneously. For example, the bivariate distribution shown in Figure 4.15, which required simultaneous estimation of 11 parameters, was fit using the EM algorithm. This problem would have been numerically impractical with the Newton–Raphson approach unless the correct answer had been known to good approximation initially.

Just what will constitute the sort of "complete" data allowing the machinery of the EM algorithm to be used smoothly will differ from problem to problem and may require some creativity to define. Accordingly, it is not practical to outline the method here in enough generality to serve as stand-alone instruction in its use, although the following example illustrates the nature of the process. Further examples of its use in the atmospheric science literature include Hannachi and O'Neill (2001), Katz and Zheng (1999), Sansom and Thomson (1992), and Smyth et al. (1999). The original source paper is Dempster et al. (1977), and the authoritative book-length treatment is McLachlan and Krishnan (1997).

Example 4.14. Fitting a Mixture of Two Gaussian Distributions with the EM Algorithm
Figure 4.14 shows a PDF fit to the Guayaquil temperature data in Table A.3, assuming a mixture distribution in the form of Equation 4.75, where both component PDFs $f_1(x)$ and $f_2(x)$ have been assumed to be Gaussian (Equation 4.24). As noted in connection with Figure 4.14, the fitting method was maximum likelihood, using the EM algorithm.

One interpretation of Equation 4.75 is that each datum x has been drawn from either $f_1(x)$ or $f_2(x)$, with overall probabilities w and $(1-w)$, respectively. It is not known which x's might have been drawn from which PDF, but if this more complete information were somehow to be available, then fitting the mixture of two Gaussian distributions indicated in Equation 4.75 would be straightforward: the parameters μ_1 and σ_1 defining the PDF $f_1(x)$ could be estimated using Equation 4.84 on the basis of the $f_1(x)$ data only, the parameters μ_2 and σ_2 defining the PDF $f_2(x)$ could be estimated using Equation 4.85 on the basis of the $f_2(x)$ data only, and the mixing parameter w could be estimated as the sample proportion of $f_1(x)$ data.

Even though the labels identifying particular x's as having been drawn from either $f_1(x)$ or $f_2(x)$ are not available (so that the data set is "incomplete"), the parameter estimation can proceed using the expected values of these hypothetical identifiers at each iteration step. If the hypothetical identifier variable would have been binary (equal to 1 for $f_1(x)$, and equal to 0 for $f_2(x)$), its expected value given each data value x_i would correspond to the probability that x_i was drawn from $f_1(x)$. The mixing parameter w would be equal to the average of these n hypothetical binary variables.

Equation 15.35 specifies the expected values of the hypothetical indicator variables (i.e., the n conditional probabilities) in terms of the two PDFs $f_1(x)$ and $f_2(x)$, and the mixing parameter w:

$$P(f_1 \mid x_i) = \frac{w f_1(x_i)}{w f_1(x_i) + (1-w) f_2(x_i)}, \quad i = 1, \ldots, n. \tag{4.88}$$

Equation 4.88 defines the E- (or expectation-) part of this implementation of the EM algorithm, where statistical expectations have been calculated for the unknown (and hypothetical) binary group membership data. Having calculated these n posterior probabilities, the updated maximum-likelihood estimate for the mixing parameter is

$$w = \frac{1}{n} \sum_{i=1}^{n} P(f_1 \mid x_i). \tag{4.89}$$

The remainder of the -M (or -maximization) part of the EM algorithm consists of ordinary maximum-likelihood estimation (Equations 4.84, for Gaussian-distribution fitting), using the expected quantities from Equation 4.88 in place of their unknown "complete-data" counterparts:

$$\hat{\mu}_1 = \frac{1}{nw} \sum_{i=1}^{n} P(f_1 \mid x_i)\, x_i, \tag{4.90a}$$

$$\hat{\mu}_2 = \frac{1}{n(1-w)} \sum_{i=1}^{n} [1 - P(f_1 \mid x_i)]\, x_i, \tag{4.90b}$$

$$\hat{\sigma}_1 = \left[\frac{1}{nw} \sum_{i=1}^{n} P(f_1 \mid x_i)(x_i - \hat{\mu}_1)^2 \right]^{1/2}, \tag{4.90c}$$

TABLE 4.7 Progress of the EM Algorithm Over the Seven Iterations Required to Fit the Mixture of Gaussian PDFs Shown in Figure 4.13.

Iteration	w	μ_1	μ_2	σ_1	σ_2	Log-Likelihood
0	0.50	22.00	28.00	1.00	1.00	−79.73
1	0.71	24.26	25.99	0.42	0.76	−22.95
2	0.73	24.28	26.09	0.43	0.72	−22.72
3	0.75	24.30	26.19	0.44	0.65	−22.42
4	0.77	24.31	26.30	0.44	0.54	−21.92
5	0.79	24.33	26.40	0.45	0.39	−21.09
6	0.80	24.34	26.47	0.46	0.27	−20.49
7	0.80	24.34	26.48	0.46	0.26	−20.48

and

$$\hat{\sigma}_2 = \left[\frac{1}{n\,(1-w)}\sum_{i=1}^{n}[1-P(f_1|\,x_i)](x_i-\hat{\mu}_2)^2\right]^{1/2}. \tag{4.90d}$$

That is, Equation 4.90 implements Equation 4.84 for each of the two Gaussian distributions $f_1(x)$ and $f_2(x)$, using expected values for the hypothetical indicator variables, rather than sorting the x's into two disjoint groups. If these hypothetical labels could be known, such a sorting would correspond to the $P(f_1|x_i)$ values being equal to the corresponding binary indicators, so that Equation 4.89 would be the relative frequency of $f_1(x)$ observations; and each x_i would contribute to either Equations 4.90a and 4.90c, or to Equations 4.90b and 4.90d, only.

This implementation of the EM algorithm, for estimating parameters of the mixture PDF for two Gaussian distributions in Equation 4.75, begins with initial guesses for the five distribution parameters μ_1, σ_1, μ_2 and σ_2, and w. These initial guesses are used in Equations 4.88 and 4.89 to obtain the initial estimates for the posterior probabilities $P(f_1|x_i)$. Updated values for the mixing parameter w and the two means and two standard deviations are then obtained using Equations 4.89 and 4.80, and the process is repeated until convergence. For many problems, including this one, it is not necessary for the initial guesses to be particularly good ones. For example, Table 4.7 outlines the progress of the EM algorithm in fitting the mixture distribution that is plotted in Figure 4.14, beginning with the rather poor initial guesses $\mu_1 = 22°C$, $\mu_2 = 28°C$, $\sigma_1 = \sigma_2 = 1°C$, and $w = 0.5$. Note that the initial guesses for the two means are not even within the range of the data (although good initial guesses would be required for complicated higher-dimensional likelihoods having multiple local maxima). Nevertheless, Table 4.7 shows that the updated means are quite near their final values after only a single iteration, and that the algorithm has converged after seven iterations. The final column in this table illustrates that the log-likelihood increases monotonically with each iteration. ◇

4.6.4. Sampling Distribution of Maximum Likelihood Estimates

Even though maximum likelihood estimates may require elaborate computations, they are still sample statistics that are functions of the underlying data. As such, they are subject to sampling variations for the

same reasons and in the same ways as more ordinary statistics, and so have sampling distributions that characterize the precision of the estimates. For sufficiently large sample sizes, these sampling distributions are approximately Gaussian, and the joint sampling distribution of simultaneously estimated parameters is approximately multivariate Gaussian (e.g., the joint sampling distribution of the estimates for α and β in Equation 4.87 is approximately bivariate normal, see Example 12.2).

Let $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_K]$ represent a K-dimensional vector of parameters to be estimated. For example in Equation 4.87, $K = 2$, $\theta_1 = \alpha$, and $\theta_2 = \beta$. The estimated variance–covariance matrix for the approximate multivariate Gaussian ($[\Sigma]$, in Equation 12.1) sampling distribution is given by the inverse of the *information matrix*,

$$Var\left(\hat{\boldsymbol{\theta}}\right) = [I(\boldsymbol{\theta})]^{-1} \tag{4.91}$$

(the matrix algebra notation is defined in Chapter 11).

The information matrix is the negative expectation of the matrix of second derivatives of the log-likelihood function with respect to the vector of parameters. In the setting of maximum likelihood the information matrix is generally estimated by the negative of the matrix $L''(\boldsymbol{\theta})$ in Equation 4.86, evaluated at the estimated parameter values $\hat{\boldsymbol{\theta}}$, which is called the *observed Fisher information*,

$$[I(\boldsymbol{\theta})] \approx - \begin{bmatrix} \partial^2 L/\partial \hat{\theta}_1^2 & \partial^2 L/\partial \hat{\theta}_1 \partial \hat{\theta}_2 & \cdots & \partial^2 L/\partial \hat{\theta}_1 \partial \hat{\theta}_K \\ \partial^2 L/\partial \hat{\theta}_2 \partial \hat{\theta}_1 & \partial^2 L/\partial \hat{\theta}_2^2 & \cdots & \partial^2 L/\partial \hat{\theta}_2 \partial \hat{\theta}_K \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 L/\partial \hat{\theta}_K \partial \hat{\theta}_1 & \partial^2 L/\partial \hat{\theta}_K \partial \hat{\theta}_2 & \cdots & \partial^2 L/\partial \hat{\theta}_K^2 \end{bmatrix}. \tag{4.92}$$

The approximation is generally close for large sample sizes. Note that the inverse of the observed Fisher information matrix appears as part of the Newton–Raphson iteration for the estimation itself, for example, for parameter estimation for the gamma distribution in Equation 4.87. One advantage of using this algorithm is therefore that the estimated variances and covariances for the joint sampling distribution of the estimated parameters will already have been calculated at the final iteration. The EM algorithm does not automatically provide these quantities, but they can, of course, be computed from the estimated parameters: either by substitution of the parameter estimates into analytical expressions for the second derivatives of the log-likelihood function, or through a finite-difference approximation to the derivatives.

4.7. STATISTICAL SIMULATION

An underlying theme of this chapter is that uncertainty in physical processes can be described by suitable probability distributions. When a component of a physical phenomenon or process of interest is uncertain, that phenomenon or process can still be studied through computer simulations, using algorithms that generate numbers that can be regarded as random samples from the relevant probability distribution(s). The generation of these apparently random numbers is called *statistical simulation*.

This section describes algorithms that are used in statistical simulation. These algorithms consist of deterministic recursive functions, so their output is not really random at all. In fact, their output can be duplicated exactly if desired, which can help in the debugging of code and is an advantage when executing controlled replication of numerical experiments. Although these algorithms are sometimes called *random-number generators*, the more correct name is *pseudo-random-number generator*, since their

deterministic output only appears to be random. However, quite useful results can be obtained by regarding them as being effectively random.

Essentially all random-number generation begins with simulation from the uniform distribution, with PDF $f(u) = 1, 0 \le u \le 1$, which was described in Section 4.7.1. Simulating values from other distributions involves transformation of one or more uniform variates. Much more on this subject than can be presented here, including computer code and pseudocode for many particular algorithms, can be found in such references as Boswell et al. (1993), Bratley et al. (1987), Dagpunar (1988), Press et al. (1986), Tezuka (1995), and the encyclopedic Devroye (1986).

The material in this section pertains to generation of scalar, independent random variates. The discussion emphasizes generation of continuous variates, but the two general methods described in Sections 4.7.2 and 4.7.3 can be used for discrete distributions as well. Extension of statistical simulation to correlated sequences is included in Sections 10.2.4 and 10.3.7 on time-domain time series models. Extensions to multivariate simulation are presented in Section 12.4.

4.7.1. Uniform Random Number Generators

As noted earlier, statistical simulation depends on the availability of a good algorithm for generating apparently random and uncorrelated samples from the uniform [0, 1] distribution, which can be transformed to simulate random sampling from other distributions. Arithmetically, uniform random number generators take an initial value of an integer, called the *seed*, operate on it to produce an updated seed value, and then rescale the updated seed to the interval [0, 1]. The initial seed value is chosen by the programmer, but usually subsequent calls to the uniform generating algorithm operate on the most recently updated seed. The arithmetic operations performed by the algorithm are fully deterministic, so restarting the generator with a previously saved seed will allow exact reproduction of the resulting "random" number sequence.

The *linear congruential generator* is the most commonly encountered algorithm for uniform random number generation, defined by

$$S_n = a\,S_{n-1} + c, \quad \text{Mod } M \tag{4.93a}$$

and

$$u_n = S_n / M. \tag{4.93b}$$

Here S_{n-1} is the seed brought forward from the previous iteration; S_n is the updated seed; and a, c, and M are integer parameters called the multiplier, increment, and modulus, respectively. The quantity u_n in Equation 4.93b is the uniform variate produced by the iteration defined by Equation 4.93. Since the updated seed S_n is the remainder when $aS_{n-1} + c$ is divided by M, S_n is necessarily smaller than M, and the quotient in Equation 4.93b will be <1. For $a > 0$ and $c \ge 0$ Equation 4.93b will be >0. The parameters in Equation 4.93a must be chosen carefully if a linear congruential generator is to work at all well. The sequence S_n repeats with a period of at most M, and it is common to choose the modulus as a prime number that is nearly as large as the largest integer that can be represented by the computer on which the algorithm will be run. Many computers use 32-bit (i.e., 4-byte) integers, and $M = 2^{31} - 1$ is a usual choice in that case, often in combination with $a = 16{,}807$ and $c = 0$.

Linear congruential generators can be adequate for some purposes, particularly in low-dimensional applications. In higher dimensions, however, their output is patterned in a way that is not space filling. In particular, pairs of successive u's from Equation 4.93b fall on a set of parallel lines in the u_n – u_{n+1} plane,

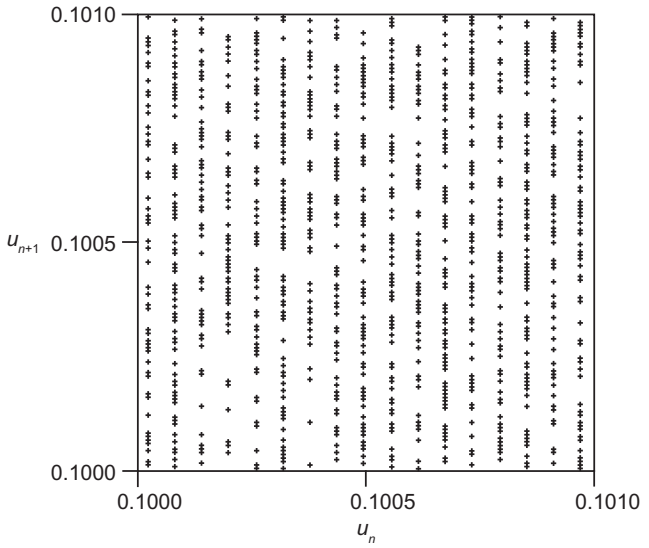

FIGURE 4.18 1000 nonoverlapping pairs of uniform random variates in a small portion of the square defined by $0 < u_n < 1$ and $0 < u_{n+1} < 1$; generated using Equation 4.93, with $a = 16{,}807$, $c = 0$, and $M = 2^{31} - 1$. This small domain contains 17 of the parallel lines onto which the successive pairs fall over the whole unit square.

triples of successive u's from Equation 4.93b fall on a set of parallel planes in the volume defined by the $u_n - u_{n+1} - u_{n+2}$ axes, and so on, with the number of these parallel features diminishing rapidly as the dimension K increases, approximately according to $(K!\ M)^{1/K}$. Here is another reason for choosing the modulus M to be as large as reasonably possible, since for $M = 2^{31} - 1$ and $K = 2$, $(K!\ M)^{1/K}$ is approximately 65,000.

Figure 4.18 shows a magnified view of a portion of the unit square, onto which 1000 nonoverlapping pairs of uniform variates generated using Equation 4.93 have been plotted. This small domain contains 17 of the parallel lines onto which successive pairs from this generator fall, which are spaced at an interval of 0.000059. Note that the minimum separation of the points in the vertical is much closer, indicating that the spacing of the nearly vertical lines of points does not define the resolution of the generator. The relatively close horizontal spacing in Figure 4.18 suggests that simple linear congruential generators may not be too crude for some low-dimensional purposes, although see Section 4.7.4 for a pathological interaction with a common algorithm for generating Gaussian variates in two dimensions. However, in higher dimensions the number of hyperplanes onto which successive groups of values from a linear congruential generator are constrained decreases rapidly, so that it is impossible for algorithms of this kind to generate many of the combinations that should be possible: for $K = 3$, 5, 10, and 20 dimensions, the number of hyperplanes containing all the supposedly randomly generated points is smaller than 2350, 200, 40, and 25, respectively, even for the relatively large modulus $M = 2^{31} - 1$. Note that the situation can be very much worse than this if the generator parameters are chosen poorly: a notorious but formerly widely used generator known as RANDU (Equation 4.93 with $a = 65{,}539$, $c = 0$, and $M = 2^{31}$) is limited to only 15 planes in three dimensions.

Direct use of linear congruential uniform generators cannot be recommended because of their patterned results in two or more dimensions. Better algorithms can be constructed by combining two or more independently running linear congruential generators, or by using one such generator to shuffle

the output of another. Examples are given in Bratley et al. (1987) and Press et al. (1986). The *Mersenne twister* (Matsumoto and Nishimura 1998), which is freely available and easily found through a Web search on that name, is an attractive alternative with apparently very good properties.

4.7.2. Nonuniform Random Number Generation by Inversion

Inversion is the easiest method of nonuniform variate generation to understand and program, when the quantile function $F^{-1}(p)$ (Equation 4.20) exists in closed form. It follows from the fact that, regardless of the functional form of the CDF $F(x)$, the distribution of the variable defined by that transformation, $u = F(x)$ follows the distribution that is uniform on [0, 1]. This relationship is called the *probability integral transform* (PIT). The converse is also true (i.e., the inverse PIT), so that the CDF of the transformed variable $x(F) = F^{-1}(u)$ is $F(x)$, if the distribution of u is uniform on [0, 1]. Therefore to generate a variate with CDF $F(x)$, for which the quantile function $F^{-1}(p)$ exists in closed form, we need only to generate a uniform variate as described in Section 4.7.1, and invert the CDF by substituting that value into the quantile function.

Inversion also can be used for distributions without closed-form quantile functions, by using numerical approximations, iterative evaluations, or interpolated table lookups. Depending on the distribution, however, these workarounds might be insufficiently fast or accurate, in which case other methods would be more appropriate.

Example 4.15. Generation of Exponential Variates Using Inversion
The exponential distribution (Equations 4.52 and 4.53) is a simple continuous distribution, for which the quantile function exists in closed form. In particular, solving Equation 4.53 for the cumulative probability p yields

$$F^{-1}(p) = -\beta \ \ln(1-p). \tag{4.94}$$

Generating exponentially distributed variates requires only that a uniform variate be substituted for the cumulative probability p in Equation 4.94, so $x(F) = F^{-1}(u) = -\beta \ln(1-u)$. Figure 4.19 illustrates the

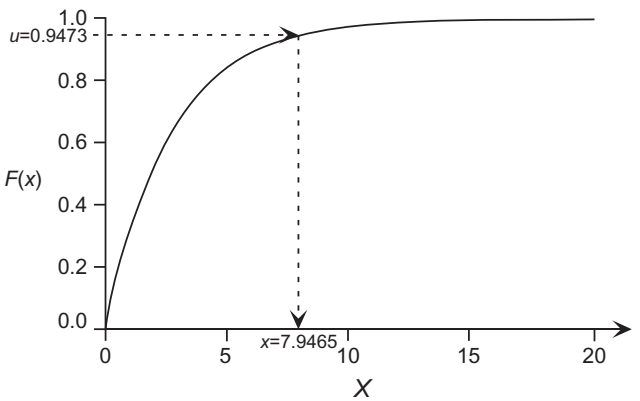

FIGURE 4.19 Illustration of the generation of an exponential variate by inversion. The smooth curve is the CDF (Equation 4.53) with mean $\beta = 2.7$. The uniform variate $u = 0.9473$ is transformed, through the inverse of the CDF, to the generated exponential variate $x = 7.9465$. This figure also illustrates that inversion produces a monotonic transformation of the underlying uniform variates.

process for an arbitrarily chosen u, and the exponential distribution with mean $\beta = 2.7$. Note that the numerical values in Figure 4.19 have been rounded to a few significant figures for convenience, but in practice all the significant digits would be retained in a computation.

Since the uniform distribution is symmetric around its middle value 0.5, the distribution of $1 - u$ is also uniform on [0, 1], so that exponential variates can be generated just as easily using $x(F) = F^{-1}(1-u) = -\beta \ln(u)$. Even though this is somewhat simpler computationally, it may be worthwhile to use $-\beta \ln(1-u)$ anyway in order to maintain the monotonicity of the inversion method. In that case the quantiles of the underlying uniform distribution correspond exactly to the quantiles of the distribution of the generated variates, so the smallest u's correspond to the smallest x's, and the largest u's correspond to the largest x's. One instance where this property can be useful is in the comparison of simulations that might depend on different parameters or different distributions. Maintaining monotonicity across such a collection of simulations (and beginning each with the same random number seed) can allow more precise comparisons among the different simulations, because a greater fraction of the variance of differences between simulations is then attributable to differences in the simulated processes, and less is due to sampling variations in the random number streams. This technique is known as *variance reduction* in the simulation literature. ◇

4.7.3. Nonuniform Random Number Generation by Rejection

The inversion method is mathematically and computationally convenient when the quantile function can be evaluated simply, but it can be awkward otherwise. The *rejection method*, or *acceptance–rejection method*, is a more general approach which requires only that the PDF, $f(x)$, of the distribution to be simulated can be evaluated explicitly. However, in addition, an envelope PDF, $g(x)$, must also be found. The envelope density $g(x)$ must have the same support as $f(x)$ and should be easy to simulate from (e.g., by inversion). In addition, a constant $c > 1$ must be found such that $f(x) \le c\ g(x)$, for all x having nonzero probability. That is, $f(x)$ must be dominated by the function $c\ g(x)$ for all relevant x. The difficult part of designing a rejection algorithm is finding an appropriate envelope PDF with a shape similar to that of the distribution to be simulated, so that the constant c can be as close to 1 as possible.

Once the envelope PDF and a constant c sufficient to ensure domination have been found, simulation by rejection proceeds in two steps, each of which requires an independent call to the uniform generator. First, a candidate variate is generated from $g(x)$ using the first uniform variate u_1, perhaps by inversion as $x = G^{-1}(u_1)$. Second, the candidate x is subjected to a random test using the second uniform variate: the candidate x is accepted if $u_2 \le f(x)/[c\ g(x)]$, otherwise the candidate x is rejected and the procedure is tried again with a new pair of uniform variates.

Figure 4.20 illustrates use of the rejection method, to simulate from the quartic density (see Table 3.1). The PDF for this distribution is a fourth-degree polynomial, so its CDF could be found easily by integration to be a fifth-degree polynomial. However, explicitly inverting the CDF (solving the fifth-degree polynomial) could be problematic, so rejection is a plausible method to simulate from this distribution. The triangular distribution (also given in Table 3.1) has been chosen as the envelope distribution $g(x)$, and the constant $c = 1.12$ is sufficient for $c\ g(x)$ to dominate $f(x)$ over $-1 \le x \le 1$. The triangular function is a reasonable choice for the envelope density because it dominates $f(x)$ with a relatively small value for the stretching constant c, so that the probability for a candidate x to be rejected is relatively small. In addition, it is simple enough that we easily can derive its quantile function, allowing simulation through inversion. In particular, integrating the triangular PDF yields the CDF

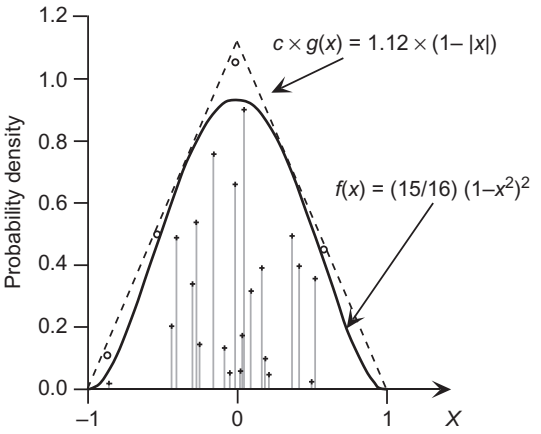

FIGURE 4.20 Illustration of simulation from the quartic (biweight) density, $f(x) = (15/16)\,(1-x^2)^2$ (Table 3.1), using a triangular density (Table 3.1) as the envelope, with $c = 1.12$. Twenty-five candidate xs have been simulated from the triangular density, of which 21 have been accepted (+) because they also fall under the distribution $f(x)$ to be simulated, and four have been rejected (○) because they fall outside it. Light gray lines point to the values simulated, on the horizontal axis.

$$G(x) = \begin{cases} \dfrac{x^2}{2} + x + \dfrac{1}{2}, & -1 \le x \le 0, \\ -\dfrac{x^2}{2} + x + \dfrac{1}{2}, & 0 \le x \le 1, \end{cases} \tag{4.95}$$

which can be inverted to obtain the quantile function

$$x(G) = G^{-1}(p) = \begin{cases} \sqrt{2p} - 1, & 0 \le p \le 1/2, \\ 1 - \sqrt{2(1-p)}, & 1/2 \le p \le 1. \end{cases} \tag{4.96}$$

Figure 4.20 indicates 25 candidate points, of which 21 have been accepted (+), with light gray lines pointing to the corresponding generated values on the horizontal axis. The horizontal coordinates of these points are $G^{-1}(u_1)$, that is, random draws from the triangular density $g(x)$ using the uniform variate u_1. Their vertical coordinates are $u_2\, c\, g[G^{-1}(u_1)]$, which is a uniformly distributed distance between the horizontal axis and $c\, g(x)$, evaluated at the candidate x using the second uniform variate u_2. Essentially, the rejection algorithm works because the two uniform variates define points distributed uniformly (in two dimensions) under the function $c\, g(x)$, and a candidate x is accepted according to the conditional probability that it is also under the PDF $f(x)$. The rejection method is thus very similar to Monte Carlo integration of $f(x)$. An illustration of simulation from this distribution by rejection is included in Example 4.16.

One drawback of the rejection method is that some pairs of uniform variates are wasted when a candidate x is rejected, and this is the reason that it is desirable for the constant c to be as small as possible: the probability that a candidate x will be rejected is $1 - 1/c$ ($= 0.107$ for the situation in Figure 4.20). Another property of the method is that an indeterminate, random number of uniform variates is required for one call to the algorithm, so that the synchronization of random number streams that is possible when using the inversion method is more difficult to achieve when using rejection.

4.7.4. Box–Muller Method for Gaussian Random Number Generation

One of the most frequently needed distributions in simulation is the Gaussian (Equation 4.24). Since the CDF for this distribution does not exist in closed form, neither does its quantile function, so generation of

Gaussian variates by inversion can be done only approximately. Alternatively, standard Gaussian (Equation 4.25) variates can be generated in pairs using a clever transformation of a pair of independent uniform variates, through an algorithm called the *Box–Muller method*. Corresponding dimensional (non-standard) Gaussian variables can then be reconstituted using the distribution mean and variance, according to Equation 4.29.

The Box–Muller method generates pairs of independent standard bivariate normal variates z_1 and z_2, that is, a random sample from the bivariate PDF in Equation 4.34, with correlation $\rho = 0$ so that the level contours of the PDF are circles. Because the level contours are circles, any direction away from the origin is equally likely, implying that in polar coordinates the PDF for the angle of a random point is uniform on $[0, 2\pi]$. A uniform angle on this interval can be easily simulated from the first of the pair of independent uniform variates as $\theta = 2\pi u_1$. The CDF for the radial distance of a standard bivariate Gaussian variate is

$$F(r) = 1 - \exp\left[-\frac{r^2}{2}\right], \quad 0 \le r \le \infty, \tag{4.97}$$

which is known as the *Rayleigh distribution*. Equation 4.97 is easily invertible to yield the quantile function $r(F) = F^{-1}(u_2) = -2\ln(1 - u_2)$. Transforming back to Cartesian coordinates, the generated pair of independent standard Gaussian variates is

$$\begin{aligned} z_1 &= \cos(2\pi u_1)\sqrt{-2\ln(u_2)}, \\ z_2 &= \sin(2\pi u_1)\sqrt{-2\ln(u_2)}. \end{aligned} \tag{4.98}$$

The Box–Muller method is very common and popular, but caution must be exercised in the choice of a uniform generator with which to drive it. In particular, the lines in the u_1–u_2 plane produced by simple linear congruential generators, illustrated in Figure 4.18, are operated upon by the polar transformation to yield spirals in the z_1–z_2 plane, as discussed in more detail by Bratley et al. (1987). This patterning is clearly undesirable, and more sophisticated uniform generators are essential when generating Box–Muller Gaussian variates.

4.7.5. Simulating from Mixture Distributions and Kernel Density Estimates

Simulation from mixture distributions (Equation 4.75) is only slightly more complicated than simulation from one of the component PDFs. It is a two-step procedure, in which a component distribution is selected according to the weights, w_i, which can be regarded as probabilities with which the component distributions will be chosen. Having randomly selected a component distribution, a variate from that distribution is generated and returned as the simulated sample from the mixture.

Consider, for example, simulation from the mixed exponential distribution, Equation 4.78, which is a probability mixture of two exponential PDFs, so that $w_1 = w$ in Equation 4.78 and $w_2 = 1 - w$. Two independent uniform variates are required in order to produce one realization from the mixture distribution: one uniform variate to choose one of the two exponential distributions, and the other to simulate from that distribution. Using inversion for the second step (Equation 4.94) the procedure is simply

$$x = \begin{cases} -\beta_1 \ln(1 - u_2), & u_1 \le w, \\ -\beta_2 \ln(1 - u_2), & u_1 > w. \end{cases} \tag{4.99}$$

Here the exponential distribution with mean β_1 is chosen with probability w, using u_1, and the inversion of whichever of the two distributions is chosen is implemented using the second uniform variate u_2.

The kernel density estimate, described in Section 3.3.6 is an interesting instance of a mixture distribution. Here the mixture consists of n equiprobable PDFs, each of which corresponds to one of n observations of a variable x, so that $w_i = 1/n$, $i = 1, \ldots, n$. These PDFs are often of one of the forms listed in Table 3.1. Again, the first step is to choose which of the n data values on which the kernel to be simulated from in the second step will be centered, which can be done according to

$$\text{choose } x_i \text{ if } \frac{i-1}{n} \le u < \frac{i}{n}, \tag{4.100a}$$

yielding

$$i = \text{int}[nu + 1]. \tag{4.100b}$$

Here int[·] indicates retention of the integer part only, or truncation of fractions.

Example 4.16. Simulation from the Kernel Density Estimate in Figure 3.8b

Figure 3.8b shows a kernel density estimate representing the Guayaquil temperature data in Table A.3, constructed using Equation 3.14, the quartic kernel (see Table 3.1), and smoothing parameter $w = 0.6$. Using rejection to simulate from the quartic kernel density, at least three independent uniform variates will be required to simulate one random sample from this distribution. Suppose these three uniform variates are generated as $u_1 = 0.257990$, $u_2 = 0.898875$, and $u_3 = 0.465617$.

The first step is to choose which of the $n = 20$ temperature values in Table A.3 will be used to center the kernel to be simulated from. Using Equation 4.100b, this will be x_i, where $i = \text{int}[20 \cdot 0.257990 + 1] = \text{int}[6.1598] = 6$, yielding $T_6 = 24.3°\text{C}$, because $i = 6$ corresponds to the year 1956 in Table A.3.

The second step is to simulate from a quartic kernel, which can be done by rejection, as illustrated in Figure 4.20. First, a candidate x is generated from the dominating triangular distribution by inversion (Equation 4.96) using the second uniform variate, $u_2 = 0.898875$. This calculation yields $x(G) = 1 - [2(1 - 0.898875)]^{1/2} = 0.550278$. Will this value be accepted or rejected? This question is answered by comparing u_3 to the ratio $f(x)/[c\, g(x)]$, where $f(x)$ is the quartic PDF, $g(x)$ is the triangular PDF, and $c = 1.12$ in order for $c\, g(x)$ to dominate $f(x)$. We find, then, that $u_3 = 0.465617 < 0.455700/[1.12 \cdot 0.449722] = 0.904726$, so the candidate $x = 0.550278$ is accepted.

The value x just generated is a random draw from a standard quartic kernel, centered on zero and having unit smoothing parameter. Equating it with the argument of the kernel function h in Equation 3.14 yields $x = 0.550278 = (T - T_6)/w = (T - 24.3°\text{C})/0.6$, which centers the kernel on T_6 and scales it appropriately, so that the final simulated value is $T = (0.550278)\ (0.6) + 24.3 = 24.63°\text{C}$. ⋄

4.8. EXERCISES

4.1. Using the binomial distribution as a model for the freezing of Cayuga Lake as presented in Examples 4.1 and 4.2, calculate the probability that the lake will freeze at least once during the four-year stay of a typical Cornell undergraduate in Ithaca.

4.2. Compute probabilities that Cayuga Lake will freeze next.
   a. In exactly 5 years.
   b. In 25 or more years.

4.3. In an article published in the journal *Science*, Gray (1990) contrasts various aspects of Atlantic hurricanes occurring in drought vs. wet years in sub-Saharan Africa. During the 18-year drought

period 1970–1987, only one strong hurricane (intensity 3 or higher) made landfall on the east coast of the United States, but 13 such storms hit the eastern United States during the 23-year wet period 1947–1969.

a. Assume that the number of hurricanes making landfall in the eastern United States follows a Poisson distribution whose characteristics depend on African rainfall. Fit two Poisson distributions to Gray's data (one conditional on drought, and one conditional on a wet year, in West Africa).
b. Compute the probability that at least one strong hurricane will hit the eastern United States, given a dry year in West Africa.
c. Compute the probability that at least one strong hurricane will hit the eastern United States, given a wet year in West Africa.

4.4. Assume that a strong hurricane making landfall in the eastern United States causes, on average, $5 billion in damage. What are the expected values of annual hurricane damage from such storms, according to each of the two conditional distributions in Exercise 4.3?

4.5. Derive a general expression for the quantity $E[X^2]$ when X is distributed according to the binomial distribution, in terms of the distribution parameters p and N.

4.6. For any Gaussian distribution,

a. What is the probability that a randomly selected data point will fall between the two quartiles?
b. What is the probability that a randomly selected data point will fall outside the inner fences (either above or below the main body of the data)?

4.7. Using the June temperature data for Guayaquil, Ecuador, in Table A.3,

a. Fit a Gaussian distribution.
b. Without converting the individual data values, determine the two Gaussian parameters that would have resulted if this data had been expressed in°F.
c. Construct a histogram of this temperature data, and superimpose the density function of the fitted distribution on the histogram plot.

4.8. Using the Gaussian distribution with $\mu = 19°C$ and $\sigma = 1.7°C$:

a. Estimate the probability that January temperature (for Miami, Florida) will be colder than 18°C.
b. What temperature will be higher than all but the warmest 1% of Januaries at Miami?

4.9. The distribution of total summer (June, July, and August) precipitation at Montpelier, Vermont, can be represented by a gamma distribution with shape parameter $\alpha = 40$ and scale parameter $\beta = 0.24$ in. Gamma distributions with shape parameters this large are well approximated by Gaussian distributions having the same mean and variance. Compare the probabilities computed using gamma and Gaussian distributions that this location will receive no more than 7 in. of precipitation in a given summer.

4.10. For the Ithaca July rainfall data given in Table 4.8,

a. Fit a gamma distribution using Thom's approximation to the MLEs.
b. Without converting the individual data values, determine the values of the two parameters that would have resulted if the data had been expressed in mm.

c. Construct a histogram of this precipitation data and superimpose the fitted gamma density function.

TABLE 4.8 July Precipitation at Ithaca, New York, 1951–1980 (inches)

1951	4.17	1961	4.24	1971	4.25
1952	5.61	1962	1.18	1972	3.66
1953	3.88	1963	3.17	1973	2.12
1954	1.55	1964	4.72	1974	1.24
1955	2.30	1965	2.17	1975	3.64
1956	5.58	1966	2.17	1976	8.44
1957	5.58	1967	3.94	1977	5.20
1958	5.14	1968	0.95	1978	2.33
1959	4.52	1969	1.48	1979	2.18
1960	1.53	1970	5.68	1980	3.43

4.11. Use the result from Exercise 4.10 to compute:
   a. The 30th and 70th percentiles of July precipitation at Ithaca.
   b. The difference between the sample mean and the median of the fitted distribution.
   c. The probability that Ithaca precipitation during any future July will be at least 7 in.

4.12. Using the lognormal distribution to represent the data in Table 4.8, recalculate Exercise 4.11.

4.13. What is the lower quartile of an Exponential distribution with $\beta = 15$ cm?

4.14. The average of the greatest snow depths for each winter at a location of interest is 80 cm, and the standard deviation (reflecting year-to-year differences in maximum snow depth) is 45 cm.
   a. Fit a Gumbel distribution to represent these data, using the method of moments.
   b. Derive the quantile function for the Gumbel distribution, and use it to estimate the snow depth that will be exceeded in only one year out of 100, on average.

4.15. Using the GEV distribution to represent the annual maximum precipitation data in Table 4.6,
   a. What is the daily precipitation amount corresponding to the 10-year return period?
   b. What is the probability that at least 2 years in the next decade will have daily precipitation amounts exceeding this value?

4.16. Consider the bivariate normal distribution as a model for the Canandaigua maximum and Canandaigua minimum temperature data in Table A.1.
   a. Fit the distribution parameters.
   b. Using the fitted distribution, compute the probability that the maximum temperature will be as cold or colder than 20°F, given that the minimum temperature is 0°F.

4.17. Construct a Q–Q plot for the temperature data in Table A.3, in comparison to the corresponding fitted Gaussian distribution.

4.18. a. Derive a formula for the MLE for the exponential distribution (Equation 4.52) parameter, β.
b. Derive a formula for the standard deviation of the sampling distribution for β, assuming n is large.

4.19. The frequency distribution for hourly averaged wind speeds at a particular location is well described by a Weibull distribution with shape parameter $\alpha = 1.2$ and scale parameter $\beta = 7.4\,\text{m/s}$. What is the probability that, during an arbitrarily selected hour, the average wind speed will be between 10 m/s and 20 m/s?

4.20. Design an algorithm to simulate from the Weibull distribution by inversion.