Chapter 3

Common Families of Distributions

"How do all these unusuals strike you, Watson?"
"Their cumulative effect is certainly considerable, and yet each of them is quite possible in itself."

Sherlock Holmes and Dr. Watson
The Adventure of the Abbey Grange

3.1 Introduction

Statistical distributions are used to model populations; as such, we usually deal with a *family* of distributions rather than a single distribution. This family is indexed by one or more parameters, which allow us to vary certain characteristics of the distribution while staying with one functional form. For example, we may specify that the normal distribution is a reasonable choice to model a particular population, but we cannot precisely specify the mean. Then, we deal with a parametric family, normal distributions with mean μ, where μ is an unspecified parameter, $-\infty < \mu < \infty$.

In this chapter we catalog many of the more common statistical distributions, some of which we have previously encountered. For each distribution we will give its mean and variance and many other useful or descriptive measures that may aid understanding. We will also indicate some typical applications of these distributions and some interesting and useful interrelationships. Some of these facts are summarized in tables at the end of the book. This chapter is by no means comprehensive in its coverage of statistical distributions. That task has been accomplished by Johnson and Kotz (1969–1972) in their multiple-volume work *Distributions in Statistics* and in the updated volumes by Johnson, Kotz, and Balakrishnan (1994, 1995) and Johnson, Kotz, and Kemp (1992).

3.2 Discrete Distributions

A random variable X is said to have a discrete distribution if the range of X, the sample space, is countable. In most situations, the random variable has integer-valued outcomes.

Discrete Uniform Distribution

A random variable X has a *discrete uniform* $(1, N)$ *distribution* if

$$P(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \ldots, N, \tag{3.2.1}$$

where N is a specified integer. This distribution puts equal mass on each of the outcomes $1, 2, \ldots, N$.

A note on notation: When we are dealing with parametric distributions, as will almost always be the case, the distribution is dependent on values of the parameters. In order to emphasize this fact and to keep track of the parameters, we write them in the pmf preceded by a "|" (given). This convention will also be used with cdfs, pdfs, expectations, and other places where it might be necessary to keep track of the parameters. When there is no possibility of confusion, the parameters may be omitted in order not to clutter up notation too much.

To calculate the mean and variance of X, recall the identities (provable by induction)

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2} \quad \text{and} \quad \sum_{i=1}^{k} i^2 = \frac{k(k+1)(2k+1)}{6}.$$

We then have

$$\mathrm{E}X = \sum_{x=1}^{N} xP(X = x|N) = \sum_{x=1}^{N} x\frac{1}{N} = \frac{N+1}{2}$$

and

$$\mathrm{E}X^2 = \sum_{x=1}^{N} x^2\frac{1}{N} = \frac{(N+1)(2N+1)}{6},$$

and so

$$\begin{aligned}\mathrm{Var}\, X &= \mathrm{E}X^2 - (\mathrm{E}X)^2 \\ &= \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 \\ &= \frac{(N+1)(N-1)}{12}.\end{aligned}$$

This distribution can be generalized so that the sample space is any range of integers, $N_0, N_0 + 1, \ldots, N_1$, with pmf $P(X = x|N_0, N_1) = 1/(N_1 - N_0 + 1)$.

Hypergeometric Distribution

The hypergeometric distribution has many applications in finite population sampling and is best understood through the classic example of the urn model.

Suppose we have a large urn filled with N balls that are identical in every way except that M are red and $N - M$ are green. We reach in, blindfolded, and select K balls at random (the K balls are taken all at once, a case of sampling without replacement). What is the probability that exactly x of the balls are red?

The total number of samples of size K that can be drawn from the N balls is $\binom{N}{K}$, as was discussed in Section 1.2.3. It is required that x of the balls be red, and this can be accomplished in $\binom{M}{x}$ ways, leaving $\binom{N-M}{K-x}$ ways of filling out the sample with $K - x$ green balls. Thus, if we let X denote the number of red balls in a sample of size K, then X has a *hypergeometric distribution* given by

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0, 1, \ldots, K. \tag{3.2.2}$$

Note that there is, implicit in (3.2.2), an additional assumption on the range of X. Binomial coefficients of the form $\binom{n}{r}$ have been defined only if $n \geq r$, and so the range of X is additionally restricted by the pair of inequalities

$$M \geq x \quad \text{and} \quad N - M \geq K - x,$$

which can be combined as

$$M - (N - K) \leq x \leq M.$$

In many cases K is small compared to M and N, so the range $0 \leq x \leq K$ will be contained in the above range and, hence, will be appropriate. The formula for the hypergeometric probability function is usually quite difficult to deal with. In fact, it is not even trivial to verify that

$$\sum_{x=0}^{K} P(X = x) = \sum_{x=0}^{K} \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} = 1.$$

The hypergeometric distribution illustrates the fact that, statistically, dealing with finite populations (finite N) is a difficult task.

The mean of the hypergeometric distribution is given by

$$\mathrm{E}X = \sum_{x=0}^{K} x \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} = \sum_{x=1}^{K} x \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}. \qquad \text{(summand is 0 at } x = 0\text{)}$$

To evaluate this expression, we use the identities (already encountered in Section 2.3)

$$x\binom{M}{x} = M\binom{M-1}{x-1},$$

$$\binom{N}{K} = \frac{N}{K}\binom{N-1}{K-1},$$

and obtain

$$\mathrm{E}X = \sum_{x=1}^{K} \frac{M\binom{M-1}{x-1}\binom{N-M}{K-x}}{\frac{N}{K}\binom{N-1}{K-1}} = \frac{KM}{N}\sum_{x=1}^{K} \frac{\binom{M-1}{x-1}\binom{N-M}{K-x}}{\binom{N-1}{K-1}}.$$

We now can recognize the second sum above as the sum of the probabilities for another hypergeometric distribution based on parameter values $N-1$, $M-1$, and $K-1$. This can be seen clearly by defining $y = x-1$ and writing

$$\sum_{x=1}^{K} \frac{\binom{M-1}{x-1}\binom{N-M}{K-x}}{\binom{N-1}{K-1}} = \sum_{y=0}^{K-1} \frac{\binom{M-1}{y}\binom{(N-1)-(M-1)}{K-1-y}}{\binom{N-1}{K-1}}$$

$$= \sum_{y=0}^{K-1} P(Y = y|N-1, M-1, K-1) = 1,$$

where Y is a hypergeometric random variable with parameters $N-1$, $M-1$, and $K-1$. Therefore, for the hypergeometric distribution,

$$\mathrm{E}X = \frac{KM}{N}.$$

A similar, but more lengthy, calculation will establish that

$$\mathrm{Var}\, X = \frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right).$$

Note the manipulations used here to calculate $\mathrm{E}X$. The sum was transformed to another hypergeometric distribution with different parameter values and, by recognizing this fact, we were able to sum the series.

Example 3.2.1 (Acceptance sampling) The hypergeometric distribution has application in acceptance sampling, as this example will illustrate. Suppose a retailer buys goods in lots and each item can be either acceptable or defective. Let

$$N = \#\text{ of items in a lot,}$$

$$M = \#\text{ of defectives in a lot.}$$

Then we can calculate the probability that a sample of size K contains x defectives. To be specific, suppose that a lot of 25 machine parts is delivered, where a part is considered acceptable only if it passes tolerance. We sample 10 parts and find that none are defective (all are within tolerance). What is the probability of this event if there are 6 defectives in the lot of 25? Applying the hypergeometric distribution with $N = 25, M = 6, K = 10$, we have

$$P(X = 0) = \frac{\binom{6}{0}\binom{19}{10}}{\binom{25}{10}} = .028,$$

showing that our observed event is quite unlikely if there are 6 (or more!) defectives in the lot. ||

Binomial Distribution

The binomial distribution, one of the more useful discrete distributions, is based on the idea of a *Bernoulli trial.* A Bernoulli trial (named for James Bernoulli, one of the founding fathers of probability theory) is an experiment with two, and only two, possible outcomes. A random variable X has a *Bernoulli(p) distribution* if

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p, \end{cases} \qquad 0 \le p \le 1. \tag{3.2.3}$$

The value $X = 1$ is often termed a "success" and p is referred to as the success probability. The value $X = 0$ is termed a "failure." The mean and variance of a Bernoulli(p) random variable are easily seen to be

$$\begin{aligned} \mathrm{E}X &= 1p + 0(1-p) = p, \\ \mathrm{Var}\, X &= (1-p)^2 p + (0-p)^2(1-p) = p(1-p). \end{aligned}$$

Many experiments can be modeled as a sequence of Bernoulli trials, the simplest being the repeated tossing of a coin; p = probability of a head, $X = 1$ if the coin shows heads. Other examples include gambling games (for example, in roulette let $X = 1$ if red occurs, so p = probability of red), election polls ($X = 1$ if candidate A gets a vote), and incidence of a disease (p = probability that a random person gets infected).

If n identical Bernoulli trials are performed, define the events

$$A_i = \{X = 1 \text{ on the } i\text{th trial}\}, \quad i = 1, 2, \ldots, n.$$

If we assume that the events $A_1, \ldots, A_n$ are a collection of independent events (as is the case in coin tossing), it is then easy to derive the distribution of the total number of successes in n trials. Define a random variable Y by

$$Y = \text{ total number of successes in } n \text{ trials.}$$

The event $\{Y = y\}$ will occur only if, out of the events $A_1, \ldots, A_n$, exactly y of them occur, and necessarily $n-y$ of them do not occur. One particular outcome (one particular ordering of occurrences and nonoccurrences) of the n Bernoulli trials might be $A_1 \cap A_2 \cap A_3^c \cap \cdots \cap A_{n-1} \cap A_n^c$. This has probability of occurrence

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3^c \cap \cdots \cap A_{n-1} \cap A_n^c) &= pp(1-p)\cdot\cdots\cdot p(1-p) \\ &= p^y(1-p)^{n-y}, \end{aligned}$$

where we have used the independence of the A_is in this calculation. Notice that the calculation is not dependent on *which* set of y A_is occurs, only that *some* set of y occurs. Furthermore, the event $\{Y = y\}$ will occur no matter which set of y A_is occurs. Putting this all together, we see that a particular sequence of n trials with exactly y successes has probability $p^y(1-p)^{n-y}$ of occurring. Since there are $\binom{n}{y}$

such sequences (the number of orderings of y 1s and $n-y$ 0s), we have

$$P(Y=y|n,p)=\binom{n}{y}p^y(1-p)^{n-y}, \quad y=0,1,2,\ldots,n,$$

and Y is called a *binomial*(n,p) *random variable.*

The random variable Y can be alternatively, and equivalently, defined in the following way: In a sequence of n identical, independent Bernoulli trials, each with success probability p, define the random variables $X_1,\ldots,X_n$ by

$$X_i=\begin{cases}1 & \text{with probability } p\\ 0 & \text{with probability } 1-p.\end{cases}$$

The random variable

$$Y=\sum_{i=1}^{n}X_i$$

has the binomial(n,p) distribution.

The fact that $\sum_{y=0}^{n}P(Y=y)=1$ follows from the following general theorem.

Theorem 3.2.2 (Binomial Theorem) *For any real numbers x and y and integer $n\geq 0$,*

$$(x+y)^n=\sum_{i=0}^{n}\binom{n}{i}x^iy^{n-i}. \tag{3.2.4}$$

Proof: Write

$$(x+y)^n=(x+y)(x+y)\cdot\cdots\cdot(x+y),$$

and consider how the right-hand side would be calculated. From each factor $(x+y)$ we choose either an x or y, and multiply together the n choices. For each $i=0,1,\ldots,n$, the number of such terms in which x appears exactly i times is $\binom{n}{i}$. Therefore, this term is of the form $\binom{n}{i}x^iy^{n-i}$ and the result follows. □

If we take $x=p$ and $y=1-p$ in (3.2.4), we get

$$1=(p+(1-p))^n=\sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i},$$

and we see that each term in the sum is a binomial probability. As another special case, take $x=y=1$ in Theorem 3.2.2 and get the identity

$$2^n=\sum_{i=0}^{n}\binom{n}{i}.$$

The mean and variance of the binomial distribution have already been derived in Examples 2.2.3 and 2.3.5, so we will not repeat the derivations here. For completeness, we state them. If $X \sim \text{binomial}(n, p)$, then

$$\text{E}X = np, \quad \text{Var}\, X = np(1-p).$$

The mgf of the binomial distribution was calculated in Example 2.3.9. It is

$$M_X(t) = [pe^t + (1-p)]^n.$$

Example 3.2.3 (Dice probabilities) Suppose we are interested in finding the probability of obtaining at least one 6 in four rolls of a fair die. This experiment can be modeled as a sequence of four Bernoulli trials with success probability $p = \frac{1}{6} = P(\text{die shows 6})$. Define the random variable X by

$$X = \text{total number of 6s in four rolls.}$$

Then $X \sim \text{binomial}\left(4, \frac{1}{6}\right)$ and

$$\begin{aligned} P(\text{at least one 6}) &= P(X > 0) = 1 - P(X = 0) \\ &= 1 - \binom{4}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^4 \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &= .518. \end{aligned}$$

Now we consider another game; throw a pair of dice 24 times and ask for the probability of at least one double 6. This, again, can be modeled by the binomial distribution with success probability p, where

$$p = P(\text{roll a double 6}) = \frac{1}{36}.$$

So, if $Y =$ number of double 6s in 24 rolls, $Y \sim \text{binomial}\left(24, \frac{1}{36}\right)$ and

$$\begin{aligned} P(\text{at least one double 6}) &= P(Y > 0) \\ &= 1 - P(Y = 0) \\ &= 1 - \binom{24}{0}\left(\frac{1}{36}\right)^0\left(\frac{35}{36}\right)^{24} \\ &= 1 - \left(\frac{35}{36}\right)^{24} \\ &= .491. \end{aligned}$$

This is the calculation originally done in the eighteenth century by Pascal at the request of the gambler de Meré, who thought both events had the same probability. (He began to believe he was wrong when he started losing money on the second bet.) ||

Poisson Distribution

The Poisson distribution is a widely applied discrete distribution and can serve as a model for a number of different types of experiments. For example, if we are modeling a phenomenon in which we are waiting for an occurrence (such as waiting for a bus, waiting for customers to arrive in a bank), the number of occurrences in a given time interval can sometimes be modeled by the Poisson distribution. One of the basic assumptions on which the Poisson distribution is built is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time. This makes it a reasonable model for situations like those indicated above. For example, it makes sense to assume that the longer we wait, the more likely it is that a customer will enter the bank. See the Miscellanea section for a more formal treatment of this.

Another area of application is in spatial distributions, where, for example, the Poisson may be used to model the distribution of bomb hits in an area or the distribution of fish in a lake.

The Poisson distribution has a single parameter λ, sometimes called the intensity parameter. A random variable X, taking values in the nonnegative integers, has a *Poisson*(λ) *distribution* if

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots. \tag{3.2.5}$$

To see that $\sum_{x=0}^{\infty} P(X = x|\lambda) = 1$, recall the Taylor series expansion of e^y,

$$e^y = \sum_{i=0}^{\infty} \frac{y^i}{i!}.$$

Thus,

$$\sum_{x=0}^{\infty} P(X = x|\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

The mean of X is easily seen to be

$$\begin{aligned}
\mathrm{E}X &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=1}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \qquad \text{(substitute } y = x - 1\text{)} \\
&= \lambda
\end{aligned}$$

A similar calculation will show that

$$\text{Var}\, X = \lambda,$$

and so the parameter λ is both the mean and the variance of the Poisson distribution.

The mgf can also be obtained by a straightforward calculation, again following from the Taylor series of e^y. We have

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

(See Exercise 2.33 and Example 2.3.13.)

Example 3.2.4 (Waiting time) As an example of a waiting-for-occurrence application, consider a telephone operator who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? At least two calls?

If we let X = number of calls in a minute, then X has a Poisson distribution with $EX = \lambda = \frac{5}{3}$. So

$$\begin{aligned} P(\text{no calls in the next minute}) &= P(X = 0) \\ &= \frac{e^{-5/3}\left(\frac{5}{3}\right)^0}{0!} \\ &= e^{-5/3} = .189; \end{aligned}$$

$$\begin{aligned} P(\text{at least two calls in the next minute}) &= P(X \geq 2) \\ &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - .189 - \frac{e^{-5/3}\left(\frac{5}{3}\right)^1}{1!} \\ &= .496. \end{aligned}$$

||

Calculation of Poisson probabilities can be done rapidly by noting the following recursion relation:

$$P(X = x) = \frac{\lambda}{x} P(X = x - 1), \quad x = 1, 2, \ldots. \tag{3.2.6}$$

This relation is easily proved by writing out the pmf of the Poisson. Similar relations hold for other discrete distributions. For example, if $Y \sim \text{binomial}(n, p)$, then

$$P(Y = y) = \frac{(n - y + 1)}{y} \frac{p}{1 - p} P(Y = y - 1). \tag{3.2.7}$$

The recursion relations (3.2.6) and (3.2.7) can be used to establish the Poisson approximation to the binomial, which we have already seen in Section 2.3, where the approximation was justified using mgfs. Set $\lambda = np$ and, if p is small, we can write

$$\frac{n - y + 1}{y} \frac{p}{1 - p} = \frac{np - p(y - 1)}{y - py} \approx \frac{\lambda}{y}$$

since, for small p, the terms $p(y-1)$ and py can be ignored. Therefore, to this level of approximation, (3.2.7) becomes

$$P(Y=y) = \frac{\lambda}{y}P(Y=y-1), \tag{3.2.8}$$

which is the Poisson recursion relation. To complete the approximation, we need only establish that $P(X=0) \approx P(Y=0)$, since all other probabilities will follow from (3.2.8). Now

$$P(Y=0) = (1-p)^n = \left(1-\frac{np}{n}\right)^n = \left(1-\frac{\lambda}{n}\right)^n$$

upon setting $np=\lambda$. Recall from Section 2.3 that for fixed λ, $\lim_{n\to\infty}(1-(\lambda/n))^n = e^{-\lambda}$, so for large n we have the approximation

$$P(Y=0) = \left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda} = P(X=0),$$

completing the Poisson approximation to the binomial.

The approximation is valid when n is large and p is small, which is exactly when it is most useful, freeing us from calculation of binomial coefficients and powers for large n.

Example 3.2.5 (Poisson approximation) A typesetter, on the average, makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

If we assume that setting a word is a Bernoulli trial with success probability $p=\frac{1}{500}$ (notice that we are labeling an error as a "success") and that the trials are independent, then X = number of errors in five pages (1500 words) is binomial$\left(1500, \frac{1}{500}\right)$. Thus

$$\begin{aligned} P(\text{no more than two errors}) &= P(X \le 2) \\ &= \sum_{x=0}^{2} \binom{1500}{x} \left(\frac{1}{500}\right)^x \left(\frac{499}{500}\right)^{1500-x} \\ &= .4230, \end{aligned}$$

which is a fairly cumbersome calculation. If we use the Poisson approximation with $\lambda = 1500\left(\frac{1}{500}\right) = 3$, we have

$$P(X \le 2) \approx e^{-3}\left(1+3+\frac{3^2}{2}\right) = .4232. \qquad \|$$

Negative Binomial Distribution

The binomial distribution counts the number of successes in a fixed number of Bernoulli trials. Suppose that, instead, we count the number of Bernoulli trials required to get a fixed number of successes. This latter formulation leads to the negative binomial distribution.

In a sequence of independent Bernoulli(p) trials, let the random variable X denote the trial at which the rth success occurs, where r is a fixed integer. Then

$$P(X = x|r,p) = \binom{x-1}{r-1} p^r(1-p)^{x-r}, \quad x = r,\ r+1,\ldots, \tag{3.2.9}$$

and we say that X has a *negative binomial*(r,p) *distribution.*

The derivation of (3.2.9) follows quickly from the binomial distribution. The event $\{X = x\}$ can occur only if there are exactly $r-1$ successes in the first $x-1$ trials, and a success on the xth trial. The probability of $r-1$ successes in $x-1$ trials is the binomial probability $\binom{x-1}{r-1} p^{r-1}(1-p)^{x-r}$, and with probability p there is a success on the xth trial. Multiplying these probabilities gives (3.2.9).

The negative binomial distribution is sometimes defined in terms of the random variable Y = number of failures before the rth success. This formulation is statistically equivalent to the one given above in terms of X = trial at which the rth success occurs, since $Y = X - r$. Using the relationship between Y and X, the alternative form of the negative binomial distribution is

$$P(Y = y) = \binom{r+y-1}{y} p^r(1-p)^y, \quad y = 0,1,\ldots. \tag{3.2.10}$$

Unless otherwise noted, when we refer to the negative binomial(r,p) distribution we will use this pmf.

The negative binomial distribution gets its name from the relationship

$$\binom{r+y-1}{y} = (-1)^y \binom{-r}{y} = (-1)^y \frac{(-r)(-r-1)(-r-2)\cdot\cdots\cdot(-r-y+1)}{(y)(y-1)(y-2)\cdot\cdots\cdot(2)(1)},$$

which is, in fact, the defining equation for binomial coefficients with negative integers (see Feller 1968 for a complete treatment). Substituting into (3.2.10) yields

$$P(Y = y) = (-1)^y \binom{-r}{y} p^r(1-p)^y,$$

which bears a striking resemblance to the binomial distribution.

The fact that $\sum_{y=0}^{\infty} P(Y = y) = 1$ is not easy to verify but follows from an extension of the Binomial Theorem, an extension that includes negative exponents. We will not pursue this further here. An excellent exposition on binomial coefficients can be found in Feller (1968).

The mean and variance of Y can be calculated using techniques similar to those used for the binomial distribution:

$$\begin{aligned} EY &= \sum_{y=0}^{\infty} y \binom{r+y-1}{y} p^r (1-p)^y \\ &= \sum_{y=1}^{\infty} \frac{(r+y-1)!}{(y-1)!(r-1)!} p^r (1-p)^y \\ &= \sum_{y=1}^{\infty} r \binom{r+y-1}{y-1} p^r (1-p)^y. \end{aligned}$$

Now write $z = y - 1$, and the sum becomes

$$\begin{aligned} EY &= \sum_{z=0}^{\infty} r \binom{r+z}{z} p^r (1-p)^{z+1} \\ &= r\frac{(1-p)}{p} \sum_{z=0}^{\infty} \binom{(r+1)+z-1}{z} p^{r+1}(1-p)^z \qquad \binom{\text{summand is negative}}{\text{binomial pmf}} \\ &= r\frac{(1-p)}{p}. \end{aligned}$$

Since the sum is over all values of a negative binomial$(r+1, p)$ distribution, it equals 1. A similar calculation will show

$$\operatorname{Var} Y = \frac{r(1-p)}{p^2}.$$

There is an interesting, and sometimes useful, reparameterization of the negative binomial distribution in terms of its mean. If we define the parameter $\mu = r(1-p)/p$, then $EY = \mu$ and a little algebra will show that

$$\operatorname{Var} Y = \mu + \frac{1}{r}\mu^2.$$

The variance is a quadratic function of the mean. This relationship can be useful in both data analysis and theoretical considerations (Morris 1982).

The negative binomial family of distributions includes the Poisson distribution as a limiting case. If $r \to \infty$ and $p \to 1$ such that $r(1-p) \to \lambda, 0 < \lambda < \infty$, then

$$\begin{aligned} EY &= \frac{r(1-p)}{p} \to \lambda, \\ \operatorname{Var} Y &= \frac{r(1-p)}{p^2} \to \lambda, \end{aligned}$$

which agree with the Poisson mean and variance. To demonstrate that the negative binomial$(r, p) \to$ Poisson(λ), we can show that all of the probabilities converge. The fact that the mgfs converge leads us to expect this (see Exercise 3.15).

Example 3.2.6 (Inverse binomial sampling) A technique known as inverse binomial sampling is useful in sampling biological populations. If the proportion of

individuals possessing a certain characteristic is p and we sample until we see r such individuals, then the number of individuals sampled is a negative binomial random variable.

For example, suppose that in a population of fruit flies we are interested in the proportion having vestigial wings and decide to sample until we have found 100 such flies. The probability that we will have to examine at least N flies is (using (3.2.9))

$$P(X \geq N) = \sum_{x=N}^{\infty} \binom{x-1}{99} p^{100}(1-p)^{x-100}$$

$$= 1 - \sum_{x=100}^{N-1} \binom{x-1}{99} p^{100}(1-p)^{x-100}.$$

For given p and N, we can evaluate this expression to determine how many fruit flies we are likely to look at. (Although the evaluation is cumbersome, the use of a recursion relation will speed things up.) ||

Example 3.2.6 shows that the negative binomial distribution can, like the Poisson, be used to model phenomena in which we are waiting for an occurrence. In the negative binomial case we are waiting for a specified number of successes.

Geometric Distribution

The geometric distribution is the simplest of the waiting time distributions and is a special case of the negative binomial distribution. If we set $r = 1$ in (3.2.9) we have

$$P(X = x|p) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots,$$

which defines the pmf of a *geometric random variable* X with success probability p. X can be interpreted as the trial at which the first success occurs, so we are "waiting for a success." The fact that $\sum_{x=1}^{\infty} P(X = x) = 1$ follows from properties of the geometric series. For any number a with $|a| < 1$,

$$\sum_{x=1}^{\infty} a^{x-1} = \frac{1}{1-a},$$

which we have already encountered in Example 1.5.4.

The mean and variance of X can be calculated by using the negative binomial formulas and by writing $X = Y + 1$ to obtain

$$\mathrm{E}X = \mathrm{E}Y + 1 = \frac{1}{p} \quad \text{and} \quad \mathrm{Var}\, X = \frac{1-p}{p^2}.$$

The geometric distribution has an interesting property, known as the "memoryless" property. For integers $s > t$, it is the case that

$$P(X > s|X > t) = P(X > s - t); \tag{3.2.11}$$

that is, the geometric distribution "forgets" what has occurred. The probability of getting an additional $s-t$ failures, having already observed t failures, is the same as the probability of observing $s-t$ failures at the start of the sequence. In other words, the probability of getting a run of failures depends only on the length of the run, not on its position.

To establish (3.2.11), we first note that for any integer n,

$$P(X > n) = P(\text{no successes in } n \text{ trials})$$

$$= (1-p)^n, \tag{3.2.12}$$

and hence

$$\begin{aligned} P(X > s | X > t) &= \frac{P(X > s \text{ and } X > t)}{P(X > t)} \\ &= \frac{P(X > s)}{P(X > t)} \\ &= (1-p)^{s-t} \\ &= P(X > s-t). \end{aligned}$$

Example 3.2.7 (Failure times) The geometric distribution is sometimes used to model "lifetimes" or "time until failure" of components. For example, if the probability is .001 that a light bulb will fail on any given day, then the probability that it will last at least 30 days is

$$P(X > 30) = \sum_{x=31}^{\infty} .001(1-.001)^{x-1} = (.999)^{30} = .970. \qquad \|$$

The memoryless property of the geometric distribution describes a very special "lack of aging" property. It indicates that the geometric distribution is not applicable to modeling lifetimes for which the probability of failure is expected to increase with time. There are other distributions used to model various types of aging; see, for example, Barlow and Proschan (1975).

3.3 Continuous Distributions

In this section we will discuss some of the more common families of continuous distributions, those with well-known names. The distributions mentioned here by no means constitute all of the distributions used in statistics. Indeed, as was seen in Section 1.6, any nonnegative, integrable function can be transformed into a pdf.

Uniform Distribution

The continuous *uniform distribution* is defined by spreading mass uniformly over an interval $[a, b]$. Its pdf is given by

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise.} \end{cases} \tag{3.3.1}$$

It is easy to check that $\int_a^b f(x)\,dx = 1$. We also have

$$\mathrm{E}X = \int_a^b \frac{x}{b-a}\,dx = \frac{b+a}{2};$$

$$\mathrm{Var}\,X = \int_a^b \frac{\left(x - \frac{b+a}{2}\right)^2}{b-a}\,dx = \frac{(b-a)^2}{12}.$$

Gamma Distribution

The gamma family of distributions is a flexible family of distributions on $[0, \infty)$ and can be derived by the construction discussed in Section 1.6. If α is a positive constant, the integral

$$\int_0^\infty t^{\alpha-1}e^{-t}\,dt$$

is finite. If α is a positive integer, the integral can be expressed in closed form; otherwise, it cannot. In either case its value defines the *gamma function*,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt. \tag{3.3.2}$$

The gamma function satisfies many useful relationships, in particular,

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \alpha > 0, \tag{3.3.3}$$

which can be verified through integration by parts. Combining (3.3.3) with the easily verified fact that $\Gamma(1) = 1$, we have for any integer $n > 0$,

$$\Gamma(n) = (n-1)!. \tag{3.3.4}$$

(Another useful special case, which will be seen in (3.3.15), is that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.)

Expressions (3.3.3) and (3.3.4) give recursion relations that ease the problems of calculating values of the gamma function. The recursion relation allows us to calculate any value of the gamma function from knowing only the values of $\Gamma(c)$, $0 < c \leq 1$.

Since the integrand in (3.3.2) is positive, it immediately follows that

$$f(t) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}, \quad 0 < t < \infty, \tag{3.3.5}$$

is a pdf. The full gamma family, however, has two parameters and can be derived by changing variables to get the pdf of the random variable $X = \beta T$ in (3.3.5), where β is a positive constant. Upon doing this, we get the *gamma*(α, β) *family*,

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}, \quad 0 < x < \infty, \quad \alpha > 0, \quad \beta > 0. \tag{3.3.6}$$

The parameter α is known as the shape parameter, since it most influences the peakedness of the distribution, while the parameter β is called the scale parameter, since most of its influence is on the spread of the distribution.

The mean of the gamma(α, β) distribution is

$$\mathrm{E}X = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x x^{\alpha-1} e^{-x/\beta}\, dx. \tag{3.3.7}$$

To evaluate (3.3.7), notice that the integrand is the kernel of a gamma$(\alpha+1, \beta)$ pdf. From (3.3.6) we know that, for any α, $\beta > 0$,

$$\int_0^\infty x^{\alpha-1} e^{-x/\beta}\, dx = \Gamma(\alpha)\beta^\alpha, \tag{3.3.8}$$

so we have

$$\begin{aligned}
\mathrm{E}X &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^\alpha e^{-x/\beta}\, dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha+1)\beta^{\alpha+1} \\
&= \frac{\alpha\Gamma(\alpha)\beta}{\Gamma(\alpha)} \qquad \text{(from (3.3.3))} \\
&= \alpha\beta.
\end{aligned}$$

Note that to evaluate $\mathrm{E}X$ we have again used the technique of recognizing the integral as the kernel of another pdf. (We have already used this technique to calculate the gamma mgf in Example 2.3.8 and, in a discrete case, to do binomial calculations in Examples 2.2.3 and 2.3.5.)

The variance of the gamma(α, β) distribution is calculated in a manner analogous to that used for the mean. In particular, in calculating $\mathrm{E}X^2$ we deal with the kernel of a gamma$(\alpha+2, \beta)$ distribution. The result is

$$\mathrm{Var}\, X = \alpha\beta^2.$$

In Example 2.3.8 we calculated the mgf of a gamma(α, β) distribution. It is given by

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha, \quad t < \frac{1}{\beta}.$$

Example 3.3.1 (Gamma-Poisson relationship) There is an interesting relationship between the gamma and Poisson distributions. If X is a gamma(α, β) random variable, where α is an integer, then for any x,

$$P(X \le x) = P(Y \ge \alpha), \tag{3.3.9}$$

where $Y \sim$ Poisson(x/β). Equation (3.3.9) can be established by successive integrations by parts, as follows. Since α is an integer, we write $\Gamma(\alpha) = (\alpha-1)!$ to get

$$\begin{aligned}
P(X \le x) &= \frac{1}{(\alpha-1)!\beta^\alpha} \int_0^x t^{\alpha-1} e^{-t/\beta}\, dt \\
&= \frac{1}{(\alpha-1)!\beta^\alpha} \left[-t^{\alpha-1}\beta e^{-t/\beta}\Big|_0^x + \int_0^x (\alpha-1)t^{\alpha-2}\beta e^{-t/\beta}\, dt \right],
\end{aligned}$$

where we use the integration by parts substitution $u = t^{\alpha-1}$, $dv = e^{-t/\beta}\,dt$. Continuing our evaluation, we have

$$P(X \le x) = \frac{-1}{(\alpha-1)!\beta^{\alpha-1}} x^{\alpha-1}e^{-x/\beta} + \frac{1}{(\alpha-2)!\beta^{\alpha-1}} \int_0^x t^{\alpha-2}e^{-t/\beta}\,dt$$
$$= \frac{1}{(\alpha-2)!\beta^{\alpha-1}} \int_0^x t^{\alpha-2}e^{-t/\beta}\,dt - P(Y = \alpha-1),$$

where $Y \sim \text{Poisson}(x/\beta)$. Continuing in this manner, we can establish (3.3.9). (See Exercise 3.19.) ||

There are a number of important special cases of the gamma distribution. If we set $\alpha = p/2$, where p is an integer, and $\beta = 2$, then the gamma pdf becomes

$$f(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1}e^{-x/2}, \quad 0 < x < \infty, \tag{3.3.10}$$

which is the *chi squared pdf with p degrees of freedom.* The mean, variance, and mgf of the chi squared distribution can all be calculated by using the previously derived gamma formulas.

The chi squared distribution plays an important role in statistical inference, especially when sampling from a normal distribution. This topic will be dealt with in detail in Chapter 5.

Another important special case of the gamma distribution is obtained when we set $\alpha = 1$. We then have

$$f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad 0 < x < \infty, \tag{3.3.11}$$

the *exponential pdf* with scale parameter β. Its mean and variance were calculated in Examples 2.2.2 and 2.3.3.

The exponential distribution can be used to model lifetimes, analogous to the use of the geometric distribution in the discrete case. In fact, the exponential distribution shares the "memoryless" property of the geometric. If $X \sim \text{exponential}(\beta)$, that is, with pdf given by (3.3.11), then for $s > t \ge 0$,

$$P(X > s|X > t) = P(X > s - t),$$

since

$$P(X > s|X > t) = \frac{P(X > s, X > t)}{P(X > t)}$$
$$= \frac{P(X > s)}{P(X > t)} \qquad (\text{since } s > t)$$
$$= \frac{\int_s^\infty \frac{1}{\beta}e^{-x/\beta}\,dx}{\int_t^\infty \frac{1}{\beta}e^{-x/\beta}\,dx}$$
$$= \frac{e^{-s/\beta}}{e^{-t/\beta}}$$

$$= e^{-(s-t)/\beta}$$
$$= P(X > s - t).$$

Another distribution related to both the exponential and the gamma families is the *Weibull distribution*. If $X \sim$ exponential(β), then $Y = X^{1/\gamma}$ has a Weibull(γ, β) distribution,

$$f_Y(y|\gamma,\beta) = \frac{\gamma}{\beta} y^{\gamma-1} e^{-y^\gamma/\beta}, \quad 0 < y < \infty, \quad \gamma > 0, \quad \beta > 0. \tag{3.3.12}$$

Clearly, we could have started with the Weibull and then derived the exponential as a special case ($\gamma = 1$). This is a matter of taste. The Weibull distribution plays an extremely important role in the analysis of failure time data (see Kalbfleisch and Prentice 1980 for a comprehensive treatment of this topic). The Weibull, in particular, is very useful for modeling *hazard functions* (see Exercises 3.25 and 3.26).

Normal Distribution

The normal distribution (sometimes called the *Gaussian distribution*) plays a central role in a large body of statistics. There are three main reasons for this. First, the normal distribution and distributions associated with it are very tractable analytically (although this may not seem so at first glance). Second, the normal distribution has the familiar bell shape, whose symmetry makes it an appealing choice for many population models. Although there are many other distributions that are also bell-shaped, most do not possess the analytic tractability of the normal. Third, there is the Central Limit Theorem (see Chapter 5 for details), which shows that, under mild conditions, the normal distribution can be used to approximate a large variety of distributions in large samples.

The normal distribution has two parameters, usually denoted by μ and σ^2, which are its mean and variance. The pdf of the *normal distribution* with mean μ and variance σ^2 (usually denoted by $\mathrm{n}(\mu, \sigma^2)$) is given by

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty. \tag{3.3.13}$$

If $X \sim \mathrm{n}(\mu, \sigma^2)$, then the random variable $Z = (X - \mu)/\sigma$ has a $\mathrm{n}(0, 1)$ distribution, also known as the *standard normal*. This is easily established by writing

$$\begin{aligned} P(Z \le z) &= P\left(\frac{X-\mu}{\sigma} \le z\right) \\ &= P(X \le z\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z\sigma+\mu} e^{-(x-\mu)^2/(2\sigma^2)}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt, \qquad \left(\text{substitute } t = \frac{x-\mu}{\sigma}\right) \end{aligned}$$

showing that $P(Z \le z)$ is the standard normal cdf.

It therefore follows that all normal probabilities can be calculated in terms of the standard normal. Furthermore, calculations of expected values can be simplified by carrying out the details in the $n(0,1)$ case, then transforming the result to the $n(\mu,\sigma^2)$ case. For example, if $Z \sim n(0,1)$,

$$EZ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-z^2/2} dz = -\frac{1}{\sqrt{2\pi}} e^{-z^2/2} \Big|_{-\infty}^{\infty} = 0,$$

and so, if $X \sim n(\mu,\sigma^2)$, it follows from Theorem 2.2.5 that

$$EX = E(\mu + \sigma Z) = \mu + \sigma EZ = \mu.$$

Similarly, we have that $\operatorname{Var} Z = 1$ and, from Theorem 2.3.4, $\operatorname{Var} X = \sigma^2$.

We have not yet established that (3.3.13) integrates to 1 over the whole real line. By applying the standardizing transformation, we need only to show that

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz = 1.$$

Notice that the integrand above is symmetric around 0, implying that the integral over $(-\infty, 0)$ is equal to the integral over $(0, \infty)$. Thus, we reduce the problem to showing

$$\int_0^{\infty} e^{-z^2/2} dz = \frac{\sqrt{2\pi}}{2} = \sqrt{\frac{\pi}{2}}. \tag{3.3.14}$$

The function $e^{-z^2/2}$ does not have an antiderivative that can be written explicitly in terms of elementary functions (that is, in closed form), so we cannot perform the integration directly. In fact, this is an example of an integration that either you know how to do or else you can spend a very long time going nowhere. Since both sides of (3.3.14) are positive, the equality will hold if we establish that the squares are equal. Square the integral in (3.3.14) to obtain

$$\begin{aligned} \left(\int_0^{\infty} e^{-z^2/2} dz \right)^2 &= \left(\int_0^{\infty} e^{-t^2/2}\, dt \right) \left(\int_0^{\infty} e^{-u^2/2} du \right) \\ &= \int_0^{\infty} \int_0^{\infty} e^{-(t^2+u^2)/2}\, dt\, du. \end{aligned}$$

The integration variables are just dummy variables, so changing their names is allowed. Now, we convert to polar coordinates. Define

$$t = r\cos\theta \quad \text{and} \quad u = r\sin\theta.$$

Then $t^2 + u^2 = r^2$ and $dt\, du = r\, d\theta\, dr$ and the limits of integration become $0 < r < \infty$, $0 < \theta < \pi/2$ (the upper limit on θ is $\pi/2$ because t and u are restricted to be positive). We now have

$$\begin{aligned}\int_0^\infty \int_0^\infty e^{-(t^2+u^2)/2}\,dt\,du &= \int_0^\infty \int_0^{\pi/2} re^{-r^2/2}\,d\theta\,dr \\ &= \frac{\pi}{2}\int_0^\infty re^{-r^2/2}dr \\ &= \frac{\pi}{2}\left[-e^{-r^2/2}\Big|_0^\infty\right] \\ &= \frac{\pi}{2},\end{aligned}$$

which establishes (3.3.14).

This integral is closely related to the gamma function; in fact, by making the substitution $w = \frac{1}{2}z^2$ in (3.3.14), we see that this integral is essentially $\Gamma(\frac{1}{2})$. If we are careful to get the constants correct, we will see that (3.3.14) implies

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty w^{-1/2}e^{-w}dw = \sqrt{\pi}. \tag{3.3.15}$$

The normal distribution is somewhat special in the sense that its two parameters, μ (the mean) and σ^2 (the variance), provide us with complete information about the exact shape and location of the distribution. This property, that the distribution is determined by μ and σ^2, is not unique to the normal pdf, but is shared by a family of pdfs called location–scale families, to be discussed in Section 3.5.

Straightforward calculus shows that the normal pdf (3.3.13) has its maximum at $x = \mu$ and inflection points (where the curve changes from concave to convex) at $\mu \pm \sigma$. Furthermore, the probability content within 1, 2, or 3 standard deviations of the mean is

$$\begin{aligned}P(|X-\mu| \le \sigma) &= P(|Z| \le 1) = .6826, \\ P(|X-\mu| \le 2\sigma) &= P(|Z| \le 2) = .9544, \\ P(|X-\mu| \le 3\sigma) &= P(|Z| \le 3) = .9974,\end{aligned}$$

where $X \sim \mathrm{n}(\mu, \sigma^2), Z \sim \mathrm{n}(0,1)$, and the numerical values can be obtained from many computer packages or from tables. Often, the two-digit values reported are .68, .95, and .99, respectively. Although these do not represent the rounded values, they are the values commonly used. Figure 3.3.1 shows the normal pdf along with these key features.

Among the many uses of the normal distribution, an important one is its use as an approximation to other distributions (which is partially justified by the Central Limit Theorem). For example, if $X \sim$ binomial(n,p), then $\mathrm{E}X = np$ and $\mathrm{Var}\,X = np(1-p)$, and under suitable conditions, the distribution of X can be approximated by that of a normal random variable with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. The "suitable conditions" are that n should be large and p should not be extreme (near 0 or 1). We want n large so that there are enough (discrete) values of X to make an approximation by a continuous distribution reasonable, and p should be "in the middle" so the binomial is nearly symmetric, as is the normal. As with most approximations there

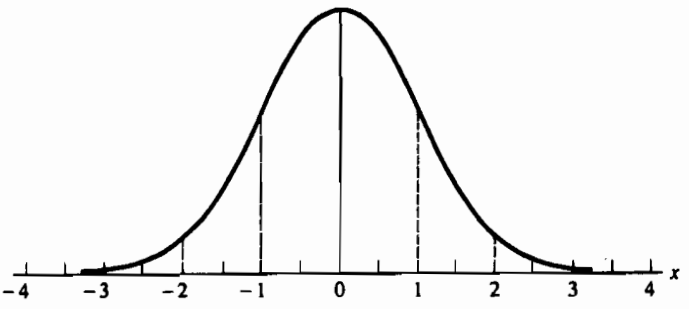

Figure 3.3.1. *Standard normal density*

are no absolute rules, and each application should be checked to decide whether the approximation is good enough for its intended use. A conservative rule to follow is that the approximation will be good if $\min(np, n(1-p)) \geq 5$.

Example 3.3.2 (Normal approximation) Let $X \sim$ binomial$(25, .6)$. We can approximate X with a normal random variable, Y, with mean $\mu = 25(.6) = 15$ and standard deviation $\sigma = ((25)(.6)(.4))^{1/2} = 2.45$. Thus

$$P(X \leq 13) \approx P(Y \leq 13) = P\left(Z \leq \frac{13-15}{2.45}\right) = P(Z \leq -.82) = .206,$$

while the exact binomial calculation gives

$$P(X \leq 13) = \sum_{x=0}^{13} \binom{25}{x} (.6)^x (.4)^{25-x} = .267,$$

showing that the normal approximation is good, but not terrific. The approximation can be greatly improved, however, by a "continuity correction." To see how this works, look at Figure 3.3.2, which shows the binomial$(25, .6)$ pmf and the n$(15, (2.45)^2)$ pdf. We have drawn the binomial pmf using bars of width 1, with height equal to the probability. Thus, the areas of the bars give the binomial probabilities. In the approximation, notice how the area of the approximating normal is smaller than the binomial area (the normal area is everything to the left of the line at 13, whereas the binomial area includes the entire bar at 13 up to 13.5). The continuity correction adds this area back by adding $\frac{1}{2}$ to the cutoff point. So instead of approximating $P(X \leq 13)$, we approximate the equivalent expression (because of the discreteness), $P(X \leq 13.5)$ and obtain

$$P(X \leq 13) = P(X \leq 13.5) \approx P(Y \leq 13.5) = P(Z \leq -.61) = .271,$$

a much better approximation. In general, the normal approximation with the continuity correction is far superior to the approximation without the continuity correction.

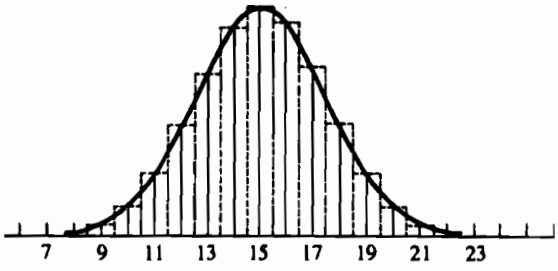

Figure 3.3.2. *Normal*$(15, (2.45)^2)$ *approximation to the binomial*$(25, .6)$

We also make the correction on the lower end. If $X \sim$ binomial(n, p) and $Y \sim$ n$(np, np(1-p))$, then we approximate

$$P(X \leq x) \approx P(Y \leq x + 1/2),$$
$$P(X \geq x) \approx P(Y \geq x - 1/2). \qquad \|$$

Beta Distribution

The beta family of distributions is a continuous family on $(0, 1)$ indexed by two parameters. The *beta*(α, β) *pdf* is

$$(3.3.16)\quad f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0,$$

where $B(\alpha, \beta)$ denotes the beta function,

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\, dx.$$

The beta function is related to the gamma function through the following identity:

$$(3.3.17)\qquad B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Equation (3.3.17) is very useful in dealing with the beta function, allowing us to take advantage of the properties of the gamma function. In fact, we will never deal directly with the beta function, but rather will use (3.3.17) for all of our evaluations.

The beta distribution is one of the few common "named" distributions that give probability 1 to a finite interval, here taken to be $(0, 1)$. As such, the beta is often used to model proportions, which naturally lie between 0 and 1. We will see illustrations of this in Chapter 4.

Calculation of moments of the beta distribution is quite easy, due to the particular form of the pdf. For $n > -\alpha$ we have

$$\begin{aligned} \mathrm{E}X^n &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^n x^{\alpha-1}(1-x)^{\beta-1}\, dx \\ &= \frac{1}{B(\alpha, \beta)} \int_0^1 x^{(\alpha+n)-1}(1-x)^{\beta-1}\, dx. \end{aligned}$$

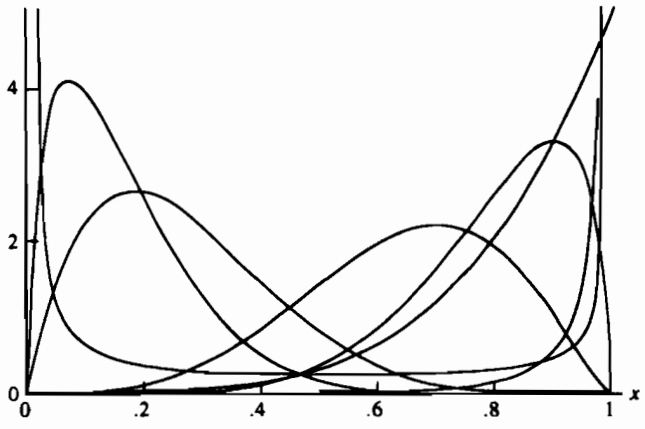

Figure 3.3.3. *Beta densities*

We now recognize the integrand as the kernel of a beta$(\alpha + n, \beta)$ pdf; hence,

$$\text{(3.3.18)} \qquad \mathrm{E}X^n = \frac{B(\alpha+n,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+n)\Gamma(\alpha)}.$$

Using (3.3.3) and (3.3.18) with $n = 1$ and $n = 2$, we calculate the mean and variance of the beta(α, β) distribution as

$$\mathrm{E}X = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \mathrm{Var}\, X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

As the parameters α and β vary, the beta distribution takes on many shapes, as shown in Figure 3.3.3. The pdf can be strictly increasing ($\alpha > 1$, $\beta = 1$), strictly decreasing ($\alpha = 1$, $\beta > 1$), U-shaped ($\alpha < 1$, $\beta < 1$), or unimodal ($\alpha > 1$, $\beta > 1$). The case $\alpha = \beta$ yields a pdf symmetric about $\frac{1}{2}$ with mean $\frac{1}{2}$ (necessarily) and variance $(4(2\alpha+1))^{-1}$. The pdf becomes more concentrated as α increases, but stays symmetric, as shown in Figure 3.3.4. Finally, if $\alpha = \beta = 1$, the beta distribution reduces to the uniform$(0, 1)$, showing that the uniform can be considered to be a member of the beta family. The beta distribution is also related, through a transformation, to the F distribution, a distribution that plays an extremely important role in statistical analysis (see Section 5.3).

Cauchy Distribution

The *Cauchy distribution* is a symmetric, bell-shaped distribution on $(-\infty, \infty)$ with pdf

$$\text{(3.3.19)} \qquad f(x|\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, \quad -\infty < x < \infty, \quad -\infty < \theta < \infty.$$

(See Exercise 3.39 for a more general version of the Cauchy pdf.) To the eye, the Cauchy does not appear very different from the normal distribution. However, there

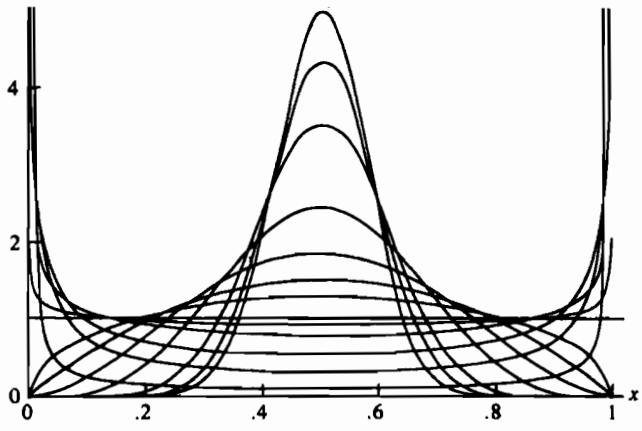

Figure 3.3.4. *Symmetric beta densities*

is a very great difference, indeed. As we have already seen in Chapter 2, the mean of the Cauchy distribution does not exist; that is,

$$\mathrm{E}|X| = \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{|x|}{1+(x-\theta)^2}\, dx = \infty. \tag{3.3.20}$$

It is easy to see that (3.3.19) defines a proper pdf for all θ. Recall that $\frac{d}{dt}\arctan(t) = (1+t^2)^{-1}$; hence,

$$\int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}\, dx = \frac{1}{\pi}\arctan(x-\theta)\Big|_{-\infty}^{\infty} = 1,$$

since $\arctan(\pm\infty) = \pm\pi/2$.

Since $\mathrm{E}|X| = \infty$, it follows that no moments of the Cauchy distribution exist or, in other words, all absolute moments equal ∞. In particular, the mgf does not exist.

The parameter θ in (3.3.19) does measure the center of the distribution; it is the median. If X has a Cauchy distribution with parameter θ, then from Exercise 3.37 it follows that $P(X \geq \theta) = \frac{1}{2}$, showing that θ is the median of the distribution. Figure 3.3.5 shows a Cauchy(0) distribution together with a n(0, 1), where we see the similarity in shape but the much thicker tails of the Cauchy.

The Cauchy distribution plays a special role in the theory of statistics. It represents an extreme case against which conjectures can be tested. But do not make the mistake of considering the Cauchy distribution to be only a pathological case, for it has a way of turning up when you least expect it. For example, it is common practice for experimenters to calculate ratios of observations, that is, ratios of random variables. (In measures of growth, it is common to combine weight and height into one measurement weight-for-height, that is, weight/height.) A surprising fact is that the ratio of two standard normals has a Cauchy distribution (see Example 4.3.6). Taking ratios can lead to ill-behaved distributions.

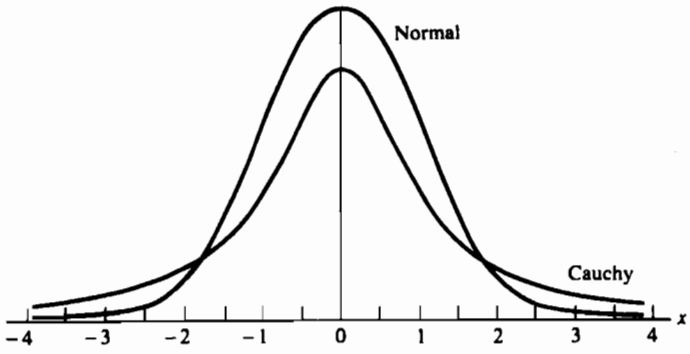

Figure 3.3.5. *Standard normal density and Cauchy density*

Lognormal Distribution

If X is a random variable whose logarithm is normally distributed (that is, $\log X \sim \mathrm{n}(\mu, \sigma^2)$), then X has a lognormal distribution. The pdf of X can be obtained by straightforward transformation of the normal pdf using Theorem 2.1.5, yielding

(3.3.21)

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-(\log x-\mu)^2/(2\sigma^2)}, \quad 0 < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0,$$

for the *lognormal pdf*. The moments of X can be calculated directly using (3.3.21), or by exploiting the relationship to the normal and writing

$$\begin{aligned} \mathrm{E}X &= \mathrm{E}e^{\log X} \\ &= \mathrm{E}e^{Y} \qquad\qquad (Y = \log X \sim \mathrm{n}(\mu,\sigma^2)) \\ &= e^{\mu+(\sigma^2/2)}. \end{aligned}$$

The last equality is obtained by recognizing the mgf of the normal distribution (set $t = 1$, see Exercise 2.33). We can use a similar technique to calculate $\mathrm{E}X^2$ and get

$$\mathrm{Var}\, X = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}.$$

The lognormal distribution is similar in appearance to the gamma distribution, as Figure 3.3.6 shows. The distribution is very popular in modeling applications when the variable of interest is skewed to the right. For example, incomes are necessarily skewed to the right, and modeling with a lognormal allows the use of normal-theory statistics on log(income), a very convenient circumstance.

Double Exponential Distribution

The *double exponential distribution* is formed by reflecting the exponential distribution around its mean. The pdf is given by

(3.3.22) $$f(x|\mu,\sigma) = \frac{1}{2\sigma}e^{-|x-\mu|/\sigma}, \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0.$$

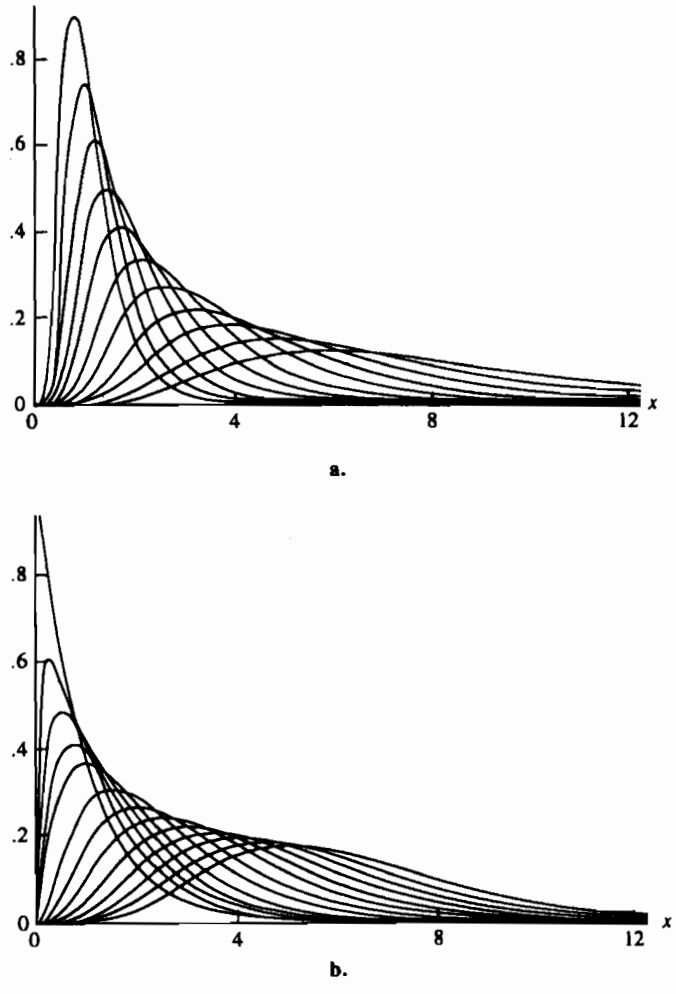

Figure 3.3.6. **(a)** *Some lognormal densities;* **(b)** *some gamma densities*

The double exponential provides a symmetric distribution with "fat" tails (much fatter than the normal) but still retains all of its moments. It is straightforward to calculate

$$\mathrm{E}X = \mu \quad \text{and} \quad \mathrm{Var}\, X = 2\sigma^2.$$

The double exponential distribution is not bell-shaped. In fact, it has a peak (or more formally, a point of nondifferentiability) at $x = \mu$. When we deal with this distribution analytically, it is important to remember this point. The absolute value signs can also be troublesome when performing integrations, and it is best to divide the integral into regions around $x = \mu$:

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} \frac{x}{2\sigma} e^{-|x-\mu|/\sigma}\, dx \\ &= \int_{-\infty}^{\mu} \frac{x}{2\sigma} e^{(x-\mu)/\sigma}\, dx + \int_{\mu}^{\infty} \frac{x}{2\sigma} e^{-(x-\mu)/\sigma}\, dx. \end{aligned} \tag{3.3.23}$$

Notice that we can remove the absolute value signs over the two regions of integration. (This strategy is useful, in general, in dealing with integrals containing absolute values; divide up the region of integration so the absolute value signs can be removed.) Evaluation of (3.3.23) can be completed by performing integration by parts on each integral.

There are many other continuous distributions that have uses in different statistical applications, many of which will appear throughout the rest of the book. The comprehensive work by Johnson and co-authors, mentioned at the beginning of this chapter, is a valuable reference for most useful statistical distributions.

3.4 Exponential Families

A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right). \tag{3.4.1}$$

Here $h(x) \geq 0$ and $t_1(x), \ldots, t_k(x)$ are real-valued functions of the observation x (they cannot depend on $\boldsymbol{\theta}$), and $c(\boldsymbol{\theta}) \geq 0$ and $w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})$ are real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta}$ (they cannot depend on x). Many common families introduced in the previous section are exponential families. These include the continuous families—normal, gamma, and beta, and the discrete families—binomial, Poisson, and negative binomial.

To verify that a family of pdfs or pmfs is an exponential family, we must identify the functions $h(x), c(\boldsymbol{\theta}), w_i(\boldsymbol{\theta})$, and $t_i(x)$ and show that the family has the form (3.4.1). The next example illustrates this.

Example 3.4.1 (Binomial exponential family) Let n be a positive integer and consider the binomial(n, p) family with $0 < p < 1$. Then the pmf for this family, for $x = 0, \ldots, n$ and $0 < p < 1$, is

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x \\ &= \binom{n}{x} (1-p)^n \exp\left(\log\left(\frac{p}{1-p}\right) x\right). \end{aligned} \tag{3.4.2}$$