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

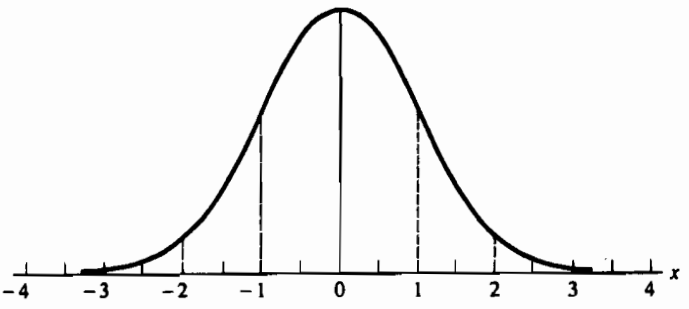

Figure 3.3.1. *Standard normal density*

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.

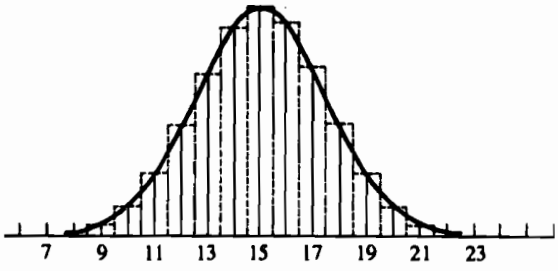

Figure 3.3.2. *Normal*$(15, (2.45)^2)$ *approximation to the binomial*$(25, .6)$

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

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

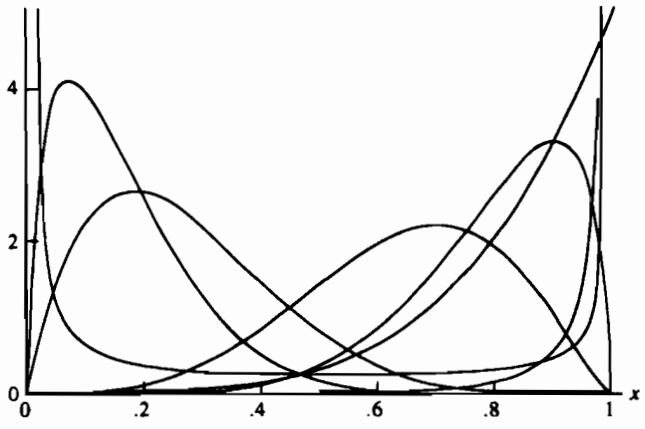

Figure 3.3.3. *Beta densities*

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

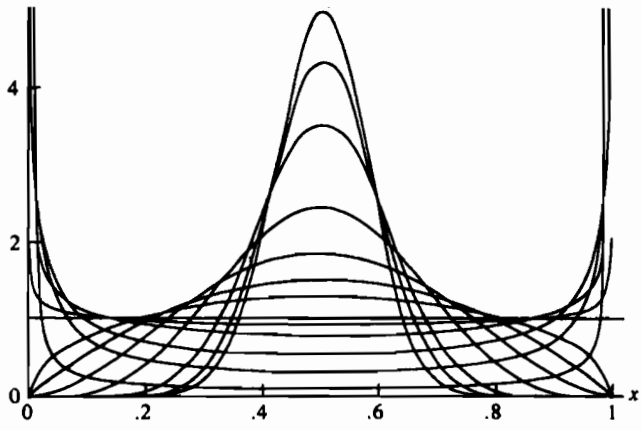

Figure 3.3.4. *Symmetric beta densities*

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.

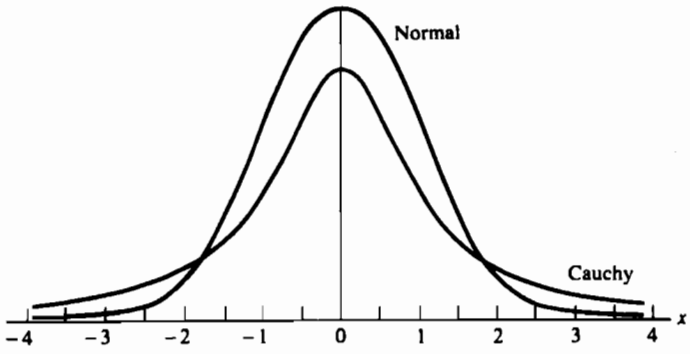

Figure 3.3.5. *Standard normal density and Cauchy density*

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

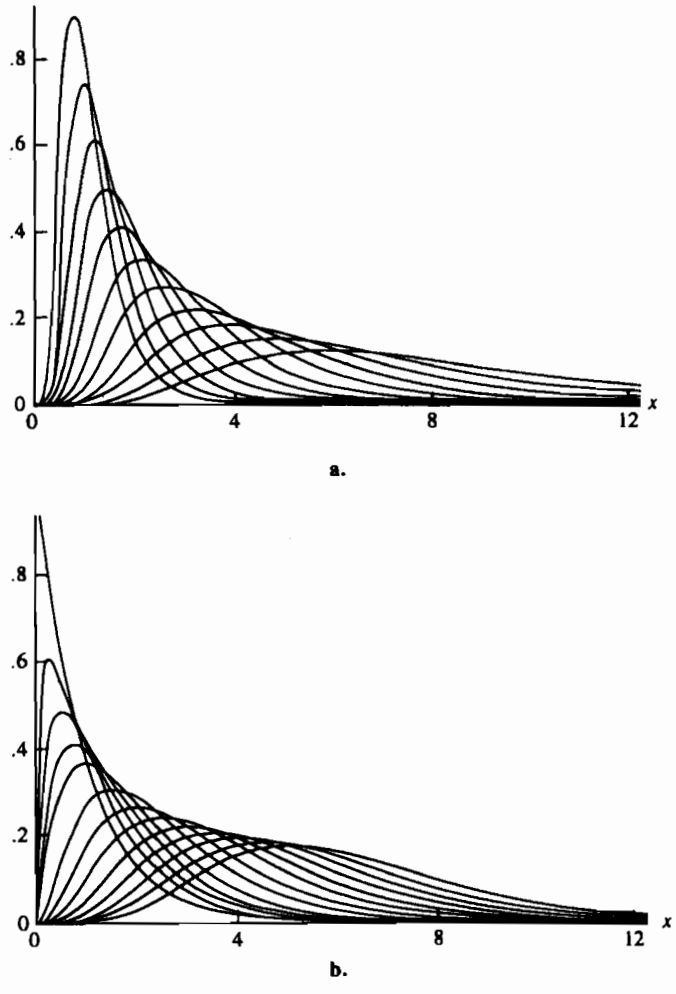

Figure 3.3.6. **(a)** *Some lognormal densities;* **(b)** *some gamma densities*

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

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

Define

$$h(x) = \begin{cases} \binom{n}{x} & x = 0, \ldots, n \\ 0 & \text{otherwise,} \end{cases} \qquad c(p) = (1-p)^n, \quad 0 < p < 1,$$

$$w_1(p) = \log\left(\frac{p}{1-p}\right), \quad 0 < p < 1, \quad \text{and} \quad t_1(x) = x.$$

Then we have

$$f(x|p) = h(x)c(p)\exp[w_1(p)t_1(x)], \tag{3.4.3}$$

which is of the form (3.4.1) with $k = 1$. In particular, note that $h(x) > 0$ only if $x = 0, \ldots, n$ and $c(p)$ is defined only if $0 < p < 1$. This is important, as (3.4.3) must match (3.4.2) for *all* values of x and is an exponential family only if $0 < p < 1$ (so the functions of the parameter are only defined here). Also, the parameter values $p = 0$ and 1 are sometimes included in the binomial model, but we have not included them here because the set of x values for which $f(x|p) > 0$ is different for $p = 0$ and 1 than for other p values. ||

The specific form of (3.4.1) results in exponential families having many nice mathematical properties. But more important for a statistical model, the form of (3.4.1) results in many nice statistical properties. We next illustrate a calculational shortcut for moments of an exponential family.

Theorem 3.4.2 *If X is a random variable with pdf or pmf of the form (3.4.1), then*

$$\mathrm{E}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial}{\partial \theta_j} \log c(\boldsymbol{\theta}); \tag{3.4.4}$$

$$\mathrm{Var}\left(\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} t_i(X)\right) = -\frac{\partial^2}{\partial \theta_j^2} \log c(\boldsymbol{\theta}) - \mathrm{E}\left(\sum_{i=1}^{k} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2} t_i(X)\right). \tag{3.4.5}$$

Although these equations may look formidable, when applied to specific cases they can work out quite nicely. Their advantage is that we can replace integration or summation by differentiation, which is often more straightforward.

Example 3.4.3 (Binomial mean and variance) From Example 3.4.1 we have

$$\frac{d}{dp} w_1(p) = \frac{d}{dp} \log \frac{p}{1-p} = \frac{1}{p(1-p)}$$

$$\frac{d}{dp} \log c(p) = \frac{d}{dp} n \log(1-p) = \frac{-n}{1-p}$$

and thus from Theorem 3.4.2 we have

$$\mathrm{E}\left(\frac{1}{p(1-p)} X\right) = \frac{n}{1-p}$$

and a bit of rearrangement yields $\mathrm{E}(X) = np$. The variance identity works in a similar manner. ||

The proof of Theorem 3.4.2 is a calculus excursion and is relegated to Exercise 3.31. See also Exercise 3.32 for a special case.

We now look at another example and some other features of exponential families.

Example 3.4.4 (Normal exponential family) Let $f(x|\mu,\sigma^2)$ be the $\mathrm{n}(\mu,\sigma^2)$ family of pdfs, where $\boldsymbol{\theta} = (\mu,\sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$. Then

$$
\begin{aligned}
f(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right).
\end{aligned} \tag{3.4.6}
$$

Define

$$
\begin{gathered}
h(x) = 1 \text{ for all } x; \\
c(\boldsymbol{\theta}) = c(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\mu^2}{2\sigma^2}\right), \quad -\infty < \mu < \infty,\ \sigma > 0; \\
w_1(\mu,\sigma) = \frac{1}{\sigma^2}, \quad \sigma > 0; \qquad w_2(\mu,\sigma) = \frac{\mu}{\sigma^2}, \ \sigma > 0; \\
t_1(x) = -x^2/2; \quad \text{and} \quad t_2(x) = x.
\end{gathered}
$$

Then

$$
f(x|\mu,\sigma^2) = h(x)c(\mu,\sigma)\exp[w_1(\mu,\sigma)t_1(x) + w_2(\mu,\sigma)t_2(x)],
$$

which is the form (3.4.1) with $k = 2$. Note again that the parameter functions are defined only over the range of the parameter. ||

In general, the set of x values for which $f(x|\boldsymbol{\theta}) > 0$ cannot depend on $\boldsymbol{\theta}$ in an exponential family. The entire definition of the pdf or pmf must be incorporated into the form (3.4.1). This is most easily accomplished by incorporating the range of x into the expression for $f(x|\boldsymbol{\theta})$ through the use of an indicator function.

Definition 3.4.5 The *indicator function* of a set A, most often denoted by $I_A(x)$, is the function

$$
I_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A. \end{cases}
$$

An alternative notation is $I(x \in A)$.

Thus, the normal pdf of Example 3.4.4 would be written

$$
f(x|\mu,\sigma^2) = h(x)c(\mu,\sigma)\exp[w_1(\mu,\sigma)t_1(x) + w_2(\mu,\sigma)t_2(x)]I_{(-\infty,\infty)}(x).
$$

Since the indicator function is a function of only x, it can be incorporated into the function $h(x)$, showing that this pdf is of the form (3.4.1).

From (3.4.1), since the factor $\exp(\cdot)$ is always positive, it can be seen that for any $\boldsymbol{\theta} \in \Theta$, that is, for any $\boldsymbol{\theta}$ for which $c(\boldsymbol{\theta}) > 0$, $\{x : f(x|\boldsymbol{\theta}) > 0\} = \{x : h(x) > 0\}$ and this set does not depend on $\boldsymbol{\theta}$. So, for example, the set of pdfs given by $f(x|\theta) = \theta^{-1}\exp(1 - (x/\theta)), 0 < \theta < x < \infty$, is not an exponential family even though we can write $\theta^{-1}\exp(1-(x/\theta)) = h(x)c(\theta)\exp(w(\theta)t(x))$, where $h(x) = e^1$, $c(\theta) = \theta^{-1}$, $w(\theta) = \theta^{-1}$, and $t(x) = -x$. Writing the pdf with indicator functions makes this very clear. We have

$$f(x|\theta) = \theta^{-1}\exp\left(1 - \left(\frac{x}{\theta}\right)\right) I_{[\theta,\infty)}(x).$$

The indicator function cannot be incorporated into any of the functions of (3.4.1) since it is not a function of x alone, not a function of θ alone, and cannot be expressed as an exponential. Thus, this is not an exponential family.

An exponential family is sometimes reparameterized as

$$f(x|\eta) = h(x)c^*(\eta)\exp\left(\sum_{i=1}^{k}\eta_i t_i(x)\right). \tag{3.4.7}$$

Here the $h(x)$ and $t_i(x)$ functions are the same as in the original parameterization (3.4.1). The set $\mathcal{H} = \left\{\eta = (\eta_1,\ldots,\eta_k) : \int_{-\infty}^{\infty} h(x)\exp\left(\sum_{i=1}^{k}\eta_i t_i(x)\right)dx < \infty\right\}$ is called the *natural parameter space* for the family. (The integral is replaced by a sum over the values of x for which $h(x) > 0$ if X is discrete.) For the values of $\eta \in \mathcal{H}$, we must have $c^*(\eta) = \left[\int_{-\infty}^{\infty} h(x)\exp\left(\sum_{i=1}^{k}\eta_i t_i(x)\right)dx\right]^{-1}$ to ensure that the pdf integrates to 1. Since the original $f(x|\boldsymbol{\theta})$ in (3.4.1) is a pdf or pmf, the set $\{\eta = (w_1(\boldsymbol{\theta}),\ldots,w_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ must be a subset of the natural parameter space. But there may be other values of $\eta \in \mathcal{H}$ also. The natural parameterization and the natural parameter space have many useful mathematical properties. For example, $\mathcal{H}$ is convex.

Example 3.4.6 (Continuation of Example 3.4.4) To determine the natural parameter space for the normal family of distributions, replace $w_i(\mu,\sigma)$ with η_i in (3.4.6) to obtain

$$f(x|\eta_1,\eta_2) = \frac{\sqrt{\eta_1}}{\sqrt{2\pi}}\exp\left(-\frac{\eta_2^2}{2\eta_1}\right)\exp\left(-\frac{\eta_1 x^2}{2} + \eta_2 x\right). \tag{3.4.8}$$

The integral will be finite if and only if the coefficient on x^2 is negative. This means η_1 must be positive. If $\eta_1 > 0$, the integral will be finite regardless of the value of η_2. Thus the natural parameter space is $\{(\eta_1,\eta_2) : \eta_1 > 0, -\infty < \eta_2 < \infty\}$. Identifying (3.4.8) with (3.4.6), we see that $\eta_2 = \mu/\sigma^2$ and $\eta_1 = 1/\sigma^2$. Although natural parameters provide a convenient mathematical formulation, they sometimes lack simple interpretations like the mean and variance. ||

In the representation (3.4.1) it is often the case that the dimension of the vector $\boldsymbol{\theta}$ is equal to k, the number of terms in the sum in the exponent. This need not be so, and it is possible for the dimension of the vector $\boldsymbol{\theta}$ to be equal to $d < k$. Such an exponential family is called a *curved exponential family.*

Definition 3.4.7 A *curved exponential family* is a family of densities of the form (3.4.1) for which the dimension of the vector $\boldsymbol{\theta}$ is equal to $d < k$. If $d = k$, the family is a *full exponential family.* (See also Miscellanea 3.8.3.)

Example 3.4.8 (A curved exponential family) The normal family of Example 3.4.4 is a full exponential family. However, if we assume that $\sigma^2 = \mu^2$, the family becomes curved. (Such a model might be used in the analysis of variance; see Exercises 11.1 and 11.2.) We then have

$$
\begin{aligned}
f(x|\mu) &= \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{(x-\mu)^2}{2\mu^2}\right) \\
&= \frac{1}{\sqrt{2\pi\mu^2}} \exp\left(-\frac{1}{2}\right) \exp\left(-\frac{x^2}{2\mu^2} + \frac{x}{\mu}\right).
\end{aligned} \tag{3.4.9}
$$

For the normal family the full exponential family would have parameter space $(\mu, \sigma^2) = \Re \times (0, \infty)$, while the parameter space of the curved family $(\mu, \sigma^2) = (\mu, \mu^2)$ is a parabola. ‖

Curved exponential families are useful in many ways. The next example illustrates a simple use.

Example 3.4.9 (Normal approximations) In Chapter 5 we will see that if $X_1, \ldots, X_n$ is a sample from a Poisson(λ) population, then the distribution of $\bar{X} = \Sigma_i X_i/n$ is approximately

$$\bar{X} \sim \mathrm{n}(\lambda, \lambda/n),$$

a curved exponential family.

The $\mathrm{n}(\lambda, \lambda/n)$ approximation is justified by the Central Limit Theorem (Theorem 5.5.14). In fact, we might realize that most such CLT approximations will result in a curved normal family. We have seen the normal binomial approximation (Example 3.3.2): If $X_1, \ldots, X_n$ are iid Bernoulli(p), then

$$\bar{X} \sim \mathrm{n}(p, p(1-p)/n),$$

approximately. For another illustration, see Example 5.5.16. ‖

Although the fact that the parameter space is a lower-dimensional space has some influence on the properties of the family, we will see that curved families still enjoy many of the properties of full families. In particular, Theorem 3.4.2 applies to curved exponential families. Moreover, full and curved exponential families have other statistical properties, which will be discussed throughout the remainder of the text. For

example, suppose we have a large number of data values from a population that has a pdf or pmf of the form (3.4.1). Then only k numbers (k = number of terms in the sum in (3.4.1)) that can be calculated from the data summarize all the information about $\boldsymbol{\theta}$ that is in the data. This "data reduction" property is treated in more detail in Chapter 6 (Theorem 6.2.10), where we discuss sufficient statistics.

For more of an introduction to exponential families, see Lehmann (1986, Section 2.7) or Lehmann and Casella (1998, Section 1.5 and Note 1.10.6). A thorough introduction, at a somewhat more advanced level, is given in the classic monograph by Brown (1986).

3.5 Location and Scale Families

In Sections 3.3 and 3.4, we discussed several common families of continuous distributions. In this section we discuss three techniques for constructing families of distributions. The resulting families have ready physical interpretations that make them useful for modeling as well as convenient mathematical properties.

The three types of families are called location families, scale families, and location–scale families. Each of the families is constructed by specifying a single pdf, say $f(x)$, called the *standard pdf* for the family. Then all other pdfs in the family are generated by transforming the standard pdf in a prescribed way. We start with a simple theorem about pdfs.

Theorem 3.5.1 *Let $f(x)$ be any pdf and let μ and $\sigma > 0$ be any given constants. Then the function*

$$g(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

is a pdf.

Proof: To verify that the transformation has produced a legitimate pdf, we need to check that $(1/\sigma)f((x-\mu)/\sigma)$, as a function of x, is a pdf for every value of μ and σ we might substitute into the formula. That is, we must check that $(1/\sigma)f((x-\mu)/\sigma)$ is nonnegative and integrates to 1. Since $f(x)$ is a pdf, $f(x) \geq 0$ for all values of x. So, $(1/\sigma)f((x-\mu)/\sigma) \geq 0$ for all values of x, μ, and σ. Next we note that

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) dx &= \int_{-\infty}^{\infty} f(y)\, dy \qquad \left(\text{substitute } y = \frac{x-\mu}{\sigma}\right) \\ &= 1, \qquad (\text{since } f(y) \text{ is a pdf})\end{aligned}$$

as was to be verified. □

We now turn to the first of our constructions, that of location families.

Definition 3.5.2 Let $f(x)$ be any pdf. Then the family of pdfs $f(x-\mu)$, indexed by the parameter μ, $-\infty < \mu < \infty$, is called the *location family with standard pdf* $f(x)$ and μ is called the *location parameter* for the family.

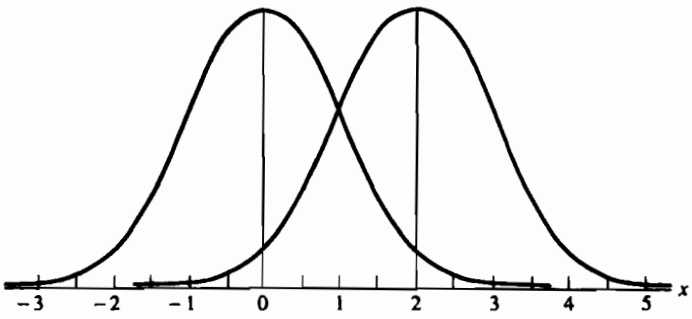

Figure 3.5.1. *Two members of the same location family: means at 0 and 2*

To see the effect of introducing the location parameter μ, consider Figure 3.5.1. At $x = \mu$, $f(x-\mu) = f(0)$; at $x = \mu+1$, $f(x-\mu) = f(1)$; and, in general, at $x = \mu + a$, $f(x-\mu) = f(a)$. Of course, $f(x-\mu)$ for $\mu = 0$ is just $f(x)$. Thus the location parameter μ simply shifts the pdf $f(x)$ so that the shape of the graph is unchanged but the point on the graph that was above $x = 0$ for $f(x)$ is above $x = \mu$ for $f(x-\mu)$. It is clear from Figure 3.5.1 that the area under the graph of $f(x)$ between $x = -1$ and $x = 2$ is the same as the area under the graph of $f(x-\mu)$ between $x = \mu - 1$ and $x = \mu + 2$. Thus if X is a random variable with pdf $f(x-\mu)$, we can write

$$P(-1 \le X \le 2|0) = P(\mu - 1 \le X \le \mu + 2|\mu),$$

where the random variable X has pdf $f(x-0) = f(x)$ on the left of the equality and pdf $f(x-\mu)$ on the right.

Several of the families introduced in Section 3.3 are, or have as subfamilies, location families. For example, if $\sigma > 0$ is a specified, known number and we define

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

then the location family with standard pdf $f(x)$ is the set of normal distributions with unknown mean μ and known variance σ^2. To see this, check that replacing x by $x-\mu$ in the above formula yields pdfs of the form defined in (3.3.13). Similarly, the Cauchy family and the double exponential family, with σ a specified value and μ a parameter, are examples of location families. But the point of Definition 3.5.2 is that we can start with *any* pdf $f(x)$ and generate a family of pdfs by introducing a location parameter.

If X is a random variable with pdf $f(x-\mu)$, then X may be represented as $X = Z + \mu$, where Z is a random variable with pdf $f(z)$. This representation is a consequence of Theorem 3.5.6 (with $\sigma = 1$), which will be proved later. Consideration of this representation indicates when a location family might be an appropriate model for an observed variable X. We will describe two such situations.

First, suppose an experiment is designed to measure some physical constant μ, say the temperature of a solution. But there is some measurement error involved in the observation. So the actual observed value X is $Z + \mu$, where Z is the measurement

error. X will be greater than μ if $Z > 0$ for this observation and less than μ if $Z < 0$. The distribution of the random measurement error might be well known from previous experience in using this measuring device to measure other solutions. If this distribution has pdf $f(z)$, then the pdf of the observed value X is $f(x - \mu)$.

As another example, suppose the distribution of reaction times of drivers on a coordination test is known from previous experimentation. Denote the reaction time for a randomly chosen driver by the random variable Z. Let the pdf of Z describing the known distribution be $f(z)$. Now, consider "applying a treatment" to the population. For example, consider what would happen if everyone drank three glasses of beer. We might assume that everyone's reaction time would change by some unknown amount μ. (This very simple model, in which everyone's reaction time changes by the same amount μ, is probably not the best model. For example, it is known that the effect of alcohol is weight-dependent, so heavier people are likely to be less affected by the beers.) Being open-minded scientists, we might even allow the possibility that $\mu < 0$, that is, that the reaction times decrease. Then, if we observe the reaction time of a randomly selected driver after "treatment," the reaction time would be $X = Z + \mu$ and the family of possible distributions for X would be given by $f(x - \mu)$.

If the set of x for which $f(x) > 0$ is not the whole real line, then the set of x for which $f(x - \mu) > 0$ will depend on μ. Example 3.5.3 illustrates this.

Example 3.5.3 (Exponential location family) Let $f(x) = e^{-x}$, $x \geq 0$, and $f(x) = 0$, $x < 0$. To form a location family we replace x with $x - \mu$ to obtain

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & x - \mu \geq 0 \\ 0 & x - \mu < 0 \end{cases}$$

$$= \begin{cases} e^{-(x-\mu)} & x \geq \mu \\ 0 & x < \mu. \end{cases}$$

Graphs of $f(x|\mu)$ for various values of μ are shown in Figure 3.5.2. As in Figure 3.5.1, the graph has been shifted. Now the positive part of the graph starts at μ rather than at 0. If X measures time, then μ might be restricted to be nonnegative so that X will be positive with probability 1 for every value of μ. In this type of model, where μ denotes a bound on the range of X, μ is sometimes called a *threshold parameter.* ||

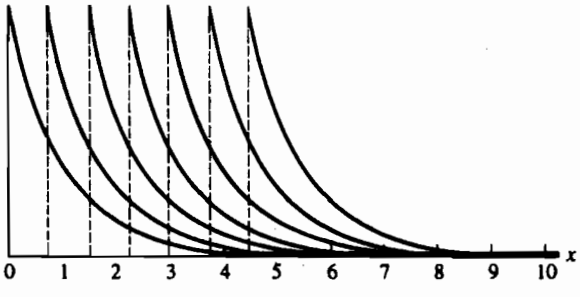

Figure 3.5.2. *Exponential location densities*

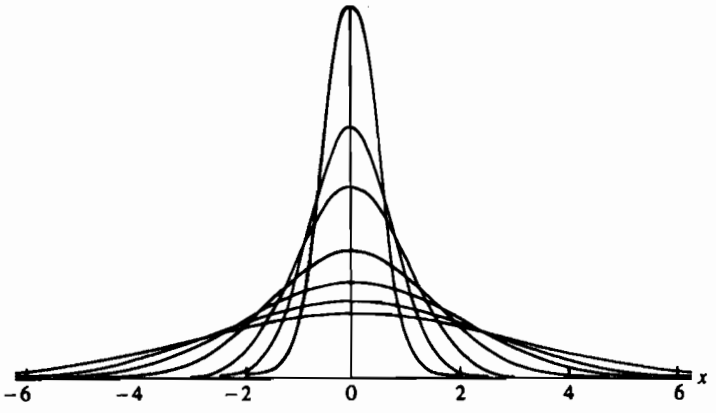

Figure 3.5.3. *Members of the same scale family*

The other two types of families to be discussed in this section are scale families and location–scale families.

Definition 3.5.4 Let $f(x)$ be any pdf. Then for any $\sigma > 0$, the family of pdfs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the *scale family with standard pdf* $f(x)$ and σ is called the *scale parameter* of the family.

The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph. This is illustrated in Figure 3.5.3. Most often when scale parameters are used, $f(x)$ is either symmetric about 0 or positive only for $x > 0$. In these cases the stretching is either symmetric about 0 or only in the positive direction. But, in the definition, any pdf may be used as the standard.

Several of the families introduced in Section 3.3 either are scale families or have scale families as subfamilies. These are the gamma family if α is a fixed value and β is the scale parameter, the normal family if $\mu = 0$ and σ is the scale parameter, the exponential family, and the double exponential family if $\mu = 0$ and σ is the scale parameter. In each case the standard pdf is the pdf obtained by setting the scale parameter equal to 1. Then all other members of the family can be shown to be of the form in Definition 3.5.4.

Definition 3.5.5 Let $f(x)$ be any pdf. Then for any μ, $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter (μ, σ), is called the *location–scale family with standard pdf* $f(x)$; μ is called the *location parameter* and σ is called the *scale parameter.*

The effect of introducing both the location and scale parameters is to stretch ($\sigma > 1$) or contract ($\sigma < 1$) the graph with the scale parameter and then shift the graph so that the point that was above 0 is now above μ. Figure 3.5.4 illustrates this transformation of $f(x)$. The normal and double exponential families are examples of location–scale families. Exercise 3.39 presents the Cauchy as a location–scale family.

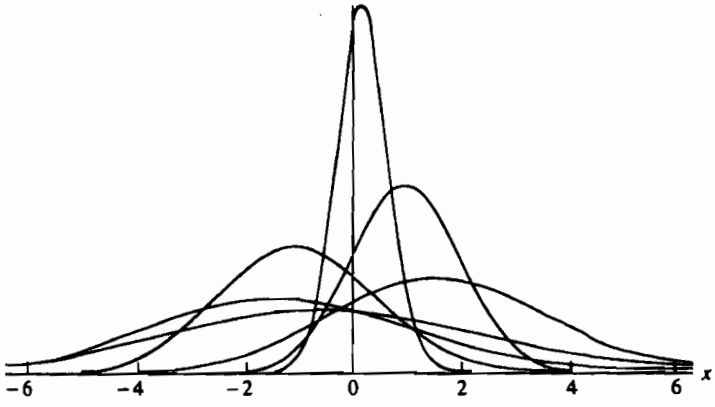

Figure 3.5.4. *Members of the same location–scale family*

The following theorem relates the transformation of the pdf $f(x)$ that defines a location–scale family to the transformation of a random variable Z with pdf $f(z)$. As mentioned earlier in the discussion of location families, the representation in terms of Z is a useful mathematical tool and can help us understand when a location–scale family might be appropriate in a modeling context. Setting $\sigma = 1$ in Theorem 3.5.6 yields a result for location (only) families, and setting $\mu = 0$ yields a result for scale (only) families.

Theorem 3.5.6 *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

Proof: To prove the "if" part, define $g(z) = \sigma z + \mu$. Then $X = g(Z)$, g is a monotone function, $g^{-1}(x) = (x-\mu)/\sigma$, and $\left|(d/dx)g^{-1}(x)\right| = 1/\sigma$. Thus by Theorem 2.1.5, the pdf of X is

$$f_X(x) = f_Z\left(g^{-1}(x)\right)\left|\frac{d}{dx}g^{-1}(x)\right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

To prove the "only if" part, define $g(x) = (x-\mu)/\sigma$ and let $Z = g(X)$. Theorem 2.1.5 again applies: $g^{-1}(z) = \sigma z + \mu$, $\left|(d/dz)g^{-1}(z)\right| = \sigma$, and the pdf of Z is

$$f_Z(z) = f_X\left(g^{-1}(z)\right)\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{1}{\sigma}f\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)\sigma = f(z).$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X. \qquad \square$$

An important fact to extract from Theorem 3.5.6 is that the random variable $Z = (X - \mu)/\sigma$ has pdf

$$f_Z(z) = \frac{1}{1} f\left(\frac{z-0}{1}\right) = f(z).$$

That is, the distribution of Z is that member of the location–scale family corresponding to $\mu = 0, \sigma = 1$. This was already proved for the special case of the normal family in Section 3.3.

Often, calculations can be carried out for the "standard" random variable Z with pdf $f(z)$ and then the corresponding result for the random variable X with pdf $(1/\sigma)f((x-\mu)/\sigma)$ can be easily derived. An example is given in the following, which is a generalization of a computation done in Section 3.3 for the normal family.

Theorem 3.5.7 *Let Z be a random variable with pdf $f(z)$. Suppose* $\mathrm{E}Z$ *and* $\mathrm{Var}\, Z$ *exist. If X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$, then*

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

In particular, if $\mathrm{E}Z = 0$ *and* $\mathrm{Var}\, Z = 1$, *then* $\mathrm{E}X = \mu$ *and* $\mathrm{Var}\, X = \sigma^2$.

Proof: By Theorem 3.5.6, there is a random variable Z^* with pdf $f(z)$ and $X = \sigma Z^* + \mu$. So $\mathrm{E}X = \sigma \mathrm{E}Z^* + \mu = \sigma \mathrm{E}Z + \mu$ and $\mathrm{Var}\, X = \sigma^2 \mathrm{Var}\, Z^* = \sigma^2 \mathrm{Var}\, Z$. □

For any location–scale family with a finite mean and variance, the standard pdf $f(z)$ can be chosen in such a way that $\mathrm{E}Z = 0$ and $\mathrm{Var}\, Z = 1$. (The proof that this choice can be made is left as Exercise 3.40.) This results in the convenient interpretation of μ and σ^2 as the mean and variance of X, respectively. This is the case for the usual definition of the normal family as given in Section 3.3. However, this is not the choice for the usual definition of the double exponential family as given in Section 3.3. There, $\mathrm{Var}\, Z = 2$.

Probabilities for any member of a location–scale family may be computed in terms of the standard variable Z because

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right).$$

Thus, if $P(Z \le z)$ is tabulated or easily calculable for the standard variable Z, then probabilities for X may be obtained. Calculations of normal probabilities using the standard normal table are examples of this.

3.6 Inequalities and Identities

Statistical theory is literally brimming with inequalities and identities—so many that entire books are devoted to the topic. The major work by Marshall and Olkin (1979) contains many inequalities using the concept of majorization. The older work by Hardy, Littlewood, and Polya (1952) is a compendium of classic inequalities. In this section and in Section 4.7 we will mix some old and some new, giving some idea of the

types of results that exist. This section is devoted to those identities and inequalities that arise from probabilistic concerns, while those in Section 4.7 rely more on basic properties of numbers and functions.

3.6.1 Probability Inequalities

The most famous, and perhaps most useful, probability inequality is Chebychev's Inequality. Its usefulness comes from its wide applicability. As with many important results, its proof is almost trivial.

Theorem 3.6.1 (Chebychev's Inequality) *Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $r > 0$,*

$$P\left(g(X) \geq r\right) \leq \frac{\mathrm{E}g(X)}{r}.$$

Proof:

$$\begin{aligned}
\mathrm{E}g(X) &= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \\
&\geq \int_{\{x:g(x)\geq r\}} g(x) f_X(x)\, dx \qquad (g \text{ is nonnegative}) \\
&\geq r \int_{\{x:g(x)\geq r\}} f_X(x)\, dx \\
&= r P\left(g(X) \geq r\right). \qquad \text{(definition)}
\end{aligned}$$

Rearranging now produces the desired inequality. □

Example 3.6.2 (Illustrating Chebychev) The most widespread use of Chebychev's Inequality involves means and variances. Let $g(x) = (x-\mu)^2/\sigma^2$, where $\mu = \mathrm{E}X$ and $\sigma^2 = \mathrm{Var}\, X$. For convenience write $r = t^2$. Then

$$P\left(\frac{(X-\mu)^2}{\sigma^2} \geq t^2\right) \leq \frac{1}{t^2}\mathrm{E}\frac{(X-\mu)^2}{\sigma^2} = \frac{1}{t^2}.$$

Doing some obvious algebra, we get the inequality

$$P(|X-\mu| \geq t\sigma) \leq \frac{1}{t^2}$$

and its companion

$$P(|X-\mu| < t\sigma) \geq 1 - \frac{1}{t^2},$$

which gives a universal bound on the deviation $|X-\mu|$ in terms of σ. For example, taking $t = 2$, we get

$$P(|X-\mu| \geq 2\sigma) \leq \frac{1}{2^2} = .25,$$

so there is at least a 75% chance that a random variable will be within 2σ of its mean (no matter what the distribution of X). ‖

While Chebychev's Inequality is widely applicable, it is necessarily conservative. (See, for example, Exercise 3.46 and Miscellanea 3.8.2.) In particular, we can often get tighter bounds for some specific distributions.

Example 3.6.3 (A normal probability inequality) If Z is standard normal, then

$$P(|Z| \geq t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}, \quad \text{for all } t > 0. \tag{3.6.1}$$

Compare this with Chebychev's Inequality. For $t = 2$, Chebychev gives $P(|Z| \geq t) \leq .25$ but $\sqrt{(2/\pi)}e^{-2}/2 = .054$, a vast improvement.

To prove (3.6.1), write

$$\begin{aligned} P(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \\ &\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2}\, dx \qquad \left(\begin{array}{c}\text{since } x/t > 1 \\ \text{for } x > t\end{array}\right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t} \end{aligned}$$

and use the fact that $P(|Z| \geq t) = 2P(Z \geq t)$. A lower bound on $P(|Z| \geq t)$ can be established in a similar way (see Exercise 3.47). ||

Many other probability inequalities exist, and almost all of them are similar in spirit to Chebychev's. For example, we will see (Exercise 3.45) that

$$P(X \geq a) \leq e^{-at} M_X(t),$$

but, of course, this inequality requires the existence of the mgf. Other inequalities, tighter than Chebychev but requiring more assumptions, exist (as detailed in Miscellanea 3.8.2).

3.6.2 Identities

In this section we present a sampling of various identities that can be useful not only in establishing theorems but also in easing numerical calculations. An entire class of identities can be thought of as "recursion relations," a few of which we have already seen. Recall that if X is Poisson(λ), then

$$P(X = x+1) = \frac{\lambda}{x+1} P(X = x), \tag{3.6.2}$$

allowing us to calculate Poisson probabilities recursively starting from $P(X = 0) = e^{-\lambda}$. Relations like (3.6.2) exist for almost all discrete distributions (see Exercise 3.48). Sometimes they exist in a slightly different form for continuous distributions.

Theorem 3.6.4 *Let $X_{\alpha,\beta}$ denote a gamma(α,β) random variable with pdf $f(x|\alpha,\beta)$, where $\alpha > 1$. Then for any constants a and b,*

$$P(a < X_{\alpha,\beta} < b) = \beta\left(f(a|\alpha,\beta) - f(b|\alpha,\beta)\right) + P(a < X_{\alpha-1,\beta} < b). \tag{3.6.3}$$

Proof: By definition,

$$\begin{aligned} P(a < X_{\alpha,\beta} < b) &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_a^b x^{\alpha-1} e^{-x/\beta}\, dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \left[-x^{\alpha-1}\beta e^{-x/\beta}\Big|_a^b + \int_a^b (\alpha-1)x^{\alpha-2}\beta e^{-x/\beta}\, dx \right], \end{aligned}$$

where we have done an integration by parts with $u = x^{\alpha-1}$ and $dv = e^{-x/\beta}\, dx$. Continuing, we have

$$P(a < X_{\alpha,\beta} < b) = \beta\left(f(a|\alpha,\beta) - f(b|\alpha,\beta)\right) + \frac{(\alpha-1)}{\Gamma(\alpha)\beta^{\alpha-1}} \int_a^b x^{\alpha-2} e^{-x/\beta}\, dx.$$

Using the fact that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, we see that the last term is $P(a < X_{\alpha-1,\beta} < b)$. □

If α is an integer, repeated use of (3.6.3) will eventually lead to an integral that can be evaluated analytically (when $\alpha = 1$, the exponential distribution). Thus, we can easily compute these gamma probabilities.

There is an entire class of identities that rely on integration by parts. The first of these is attributed to Charles Stein, who used it in his work on estimation of multivariate normal means (Stein 1973, 1981).

Lemma 3.6.5 (Stein's Lemma) *Let $X \sim \mathrm{n}(\theta,\sigma^2)$, and let g be a differentiable function satisfying $\mathrm{E}\,|g'(X)| < \infty$. Then*

$$\mathrm{E}\left[g(X)(X-\theta)\right] = \sigma^2 \mathrm{E} g'(X).$$

Proof: The left-hand side is

$$\mathrm{E}\left[g(X)(X-\theta)\right] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} g(x)(x-\theta) e^{-(x-\theta)^2/(2\sigma^2)}\, dx.$$

Use integration by parts with $u = g(x)$ and $dv = (x-\theta)e^{-(x-\theta)^2/2\sigma^2}\, dx$ to get

$$\mathrm{E}\left[g(X)(X-\theta)\right] = \frac{1}{\sqrt{2\pi}\sigma} \left[-\sigma^2 g(x) e^{-(x-\theta)^2/(2\sigma^2)}\Big|_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} g'(x) e^{-(x-\theta)^2/(2\sigma^2)}\, dx \right].$$

The condition on g' is enough to ensure that the first term is 0 and what remains on the right-hand side is $\sigma^2 \mathrm{E} g'(X)$. □

Example 3.6.6 (Higher-order normal moments) Stein's Lemma makes calculation of higher-order moments quite easy. For example, if $X \sim \mathrm{n}(\theta, \sigma^2)$, then

$$\begin{aligned} \mathrm{E}X^3 &= \mathrm{E}X^2(X-\theta+\theta) \\ &= \mathrm{E}X^2(X-\theta) + \theta\mathrm{E}X^2 \\ &= 2\sigma^2\mathrm{E}X + \theta\mathrm{E}X^2 \qquad (g(x)=x^2,\ g'(x)=2x) \\ &= 2\sigma^2\theta + \theta(\sigma^2+\theta^2) \\ &= 3\theta\sigma^2 + \theta^3. \end{aligned}$$

‖

Similar integration-by-parts identities exist for many distributions (see Exercise 3.49 and Hudson 1978). One can also get useful identities by exploiting properties of a particular distribution, as the next theorem shows.

Theorem 3.6.7 *Let χ^2_p denote a chi squared random variable with p degrees of freedom. For any function $h(x)$,*

$$\mathrm{E}h(\chi^2_p) = p\mathrm{E}\left(\frac{h\left(\chi^2_{p+2}\right)}{\chi^2_{p+2}}\right) \tag{3.6.4}$$

provided the expectations exist.

Proof: The phrase "provided the expectations exist" is a lazy way of avoiding specification of conditions on h. In general, reasonable functions will satisfy (3.6.4). We have

$$\begin{aligned} \mathrm{E}h(\chi^2_p) &= \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^\infty h(x) x^{(p/2)-1} e^{-x/2}\, dx \\ &= \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^\infty \left(\frac{h(x)}{x}\right) x^{((p+2)/2)-1} e^{-x/2}\, dx, \end{aligned}$$

where we have multiplied the integrand by x/x. Now write

$$\Gamma\left(\frac{p}{2}\right) 2^{p/2} = \frac{\Gamma((p+2)/2)2^{(p+2)/2}}{p},$$

so we have

$$\begin{aligned} \mathrm{E}h(\chi^2_p) &= \frac{p}{\Gamma((p+2)/2)2^{(p+2)/2}} \int_0^\infty \left(\frac{h(x)}{x}\right) x^{((p+2)/2)-1} e^{-x/2}\, dx \\ &= p\mathrm{E}\left(\frac{h(\chi^2_{p+2})}{\chi^2_{p+2}}\right). \end{aligned}$$

□

Some moment calculations are very easy with (3.6.4). For example, the mean of a χ^2_p is

$$\mathrm{E}\chi^2_p = p\mathrm{E}\left(\frac{\chi^2_{p+2}}{\chi^2_{p+2}}\right) = p\mathrm{E}(1) = p,$$

and the second moment is

$$\mathrm{E}(\chi^2_p)^2 = p\mathrm{E}\left(\frac{\left(\chi^2_{p+2}\right)^2}{\chi^2_{p+2}}\right) = p\mathrm{E}\left(\chi^2_{p+2}\right) = p(p+2).$$

So $\mathrm{Var}\,\chi^2_p = p(p+2) - p^2 = 2p$.

We close our section on identities with some discrete analogs of the previous identities. A general version of the two identities in Theorem 3.6.8 is due to Hwang (1982).

Theorem 3.6.8 (Hwang) *Let $g(x)$ be a function with $-\infty < \mathrm{E}g(X) < \infty$ and $-\infty < g(-1) < \infty$. Then:*

a. *If $X \sim$ Poisson(λ),*

$$\mathrm{E}\left(\lambda g(X)\right) = \mathrm{E}\left(Xg(X-1)\right). \tag{3.6.5}$$

b. *If $X \sim$ negative binomial(r, p),*

$$\mathrm{E}\left((1-p)g(X)\right) = \mathrm{E}\left(\frac{X}{r+X-1}g(X-1)\right). \tag{3.6.6}$$

Proof: We will prove part (a), saving part (b) for Exercise 3.50. We have

$$\begin{aligned}
\mathrm{E}\left(\lambda g(X)\right) &= \sum_{x=0}^{\infty} \lambda g(x) \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} g(x) \frac{e^{-\lambda}\lambda^{x+1}}{x!} \frac{(x+1)}{(x+1)} \\
&= \sum_{x=0}^{\infty} (x+1) g(x) \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}.
\end{aligned}$$

Now transform the summation index, writing $y = x+1$. As x goes from 0 to ∞, y goes from 1 to ∞. Thus

$$\begin{aligned}
\mathrm{E}\left(\lambda g(X)\right) &= \sum_{y=1}^{\infty} y g(y-1) \frac{e^{-\lambda}\lambda^y}{y!} \\
&= \sum_{y=0}^{\infty} y g(y-1) \frac{e^{-\lambda}\lambda^y}{y!} \qquad \text{(added term is 0)} \\
&= \mathrm{E}\left(Xg(X-1)\right),
\end{aligned}$$

since this last sum is a Poisson(λ) expectation. □

Hwang (1982) used his identity in a manner similar to Stein, proving results about multivariate estimators. The identity has other applications, in particular in moment calculations.

Example 3.6.9 (Higher-order Poisson moments) For $X \sim \text{Poisson}(\lambda)$, take $g(x) = x^2$ and use (3.6.5):

$$\mathrm{E}(\lambda X^2) = \mathrm{E}\left(X(X-1)^2\right) = \mathrm{E}(X^3 - 2X^2 + X).$$

Therefore, the third moment of a Poisson(λ) is

$$\begin{aligned}\mathrm{E}X^3 &= \lambda \mathrm{E}X^2 + 2\mathrm{E}X^2 - \mathrm{E}X \\ &= \lambda(\lambda + \lambda^2) + 2(\lambda + \lambda^2) - \lambda \\ &= \lambda^3 + 3\lambda^2 + \lambda.\end{aligned}$$

For the negative binomial, the mean can be calculated by taking $g(x) = r + x$ in (3.6.6):

$$\mathrm{E}\left((1-p)(r+X)\right) = \mathrm{E}\left(\frac{X}{r+X-1}(r+X-1)\right) = \mathrm{E}X,$$

so, rearranging, we get

$$(\mathrm{E}X)\left((1-p)-1\right) = -r(1-p)$$

or

$$\mathrm{E}X = \frac{r(1-p)}{p}.$$

Other moments can be calculated similarly. ‖

3.7 Exercises

3.1 Find expressions for $\mathrm{E}X$ and $\mathrm{Var}\,X$ if X is a random variable with the general discrete uniform(N_0, N_1) distribution that puts equal probability on each of the values $N_0, N_0+1, \ldots, N_1$. Here $N_0 \le N_1$ and both are integers.

3.2 A manufacturer receives a lot of 100 parts from a vendor. The lot will be unacceptable if more than five of the parts are defective. The manufacturer is going to select randomly K parts from the lot for inspection and the lot will be accepted if no defective parts are found in the sample.

(a) How large does K have to be to ensure that the probability that the manufacturer accepts an unacceptable lot is less than .10?

(b) Suppose the manufacturer decides to accept the lot if there is at most one defective in the sample. How large does K have to be to ensure that the probability that the manufacturer accepts an unacceptable lot is less than .10?

3.3 The flow of traffic at certain street corners can sometimes be modeled as a sequence of Bernoulli trials by assuming that the probability of a car passing during any given second is a constant p and that there is no interaction between the passing of cars at different seconds. If we treat seconds as indivisible time units (trials), the Bernoulli model applies. Suppose a pedestrian can cross the street only if no car is to pass during the next 3 seconds. Find the probability that the pedestrian has to wait for exactly 4 seconds before starting to cross.

3.4 A man with n keys wants to open his door and tries the keys at random. Exactly one key will open the door. Find the mean number of trials if

(a) unsuccessful keys are not eliminated from further selections.

(b) unsuccessful keys are eliminated.

3.5 A standard drug is known to be effective in 80% of the cases in which it is used. A new drug is tested on 100 patients and found to be effective in 85 cases. Is the new drug superior? (*Hint:* Evaluate the probability of observing 85 or more successes assuming that the new and old drugs are equally effective.)

3.6 A large number of insects are expected to be attracted to a certain variety of rose plant. A commercial insecticide is advertised as being 99% effective. Suppose 2,000 insects infest a rose garden where the insecticide has been applied, and let X = number of surviving insects.

(a) What probability distribution might provide a reasonable model for this experiment?

(b) Write down, but do not evaluate, an expression for the probability that fewer than 100 insects survive, using the model in part (a).

(c) Evaluate an approximation to the probability in part (b).

3.7 Let the number of chocolate chips in a certain type of cookie have a Poisson distribution. We want the probability that a randomly chosen cookie has at least two chocolate chips to be greater than .99. Find the smallest value of the mean of the distribution that ensures this probability.

3.8 Two movie theaters compete for the business of 1,000 customers. Assume that each customer chooses between the movie theaters independently and with "indifference." Let N denote the number of seats in each theater.

(a) Using a binomial model, find an expression for N that will guarantee that the probability of turning away a customer (because of a full house) is less than 1%.

(b) Use the normal approximation to get a numerical value for N.

3.9 Often, news stories that are reported as startling "one-in-a-million" coincidences are actually, upon closer examination, not rare events and can even be expected to occur. A few years ago an elementary school in New York state reported that its incoming kindergarten class contained five sets of twins. This, of course, was reported throughout the state, with a quote from the principal that this was a "statistical impossibility". Was it? Or was it an instance of what Diaconis and Mosteller (1989) call the "law of truly large numbers"? Let's do some calculations.

(a) The probability of a twin birth is approximately 1/90, and we can assume that an elementary school will have approximately 60 children entering kindergarten (three classes of 20 each). Explain how our "statistically impossible" event can be thought of as the probability of 5 or more successes from a binomial(60, 1/90). Is this even rare enough to be newsworthy?

(b) Even if the probability in part (a) is rare enough to be newsworthy, consider that this could have happened in any school in the county, and in any county in the state, and it still would have been reported exactly the same. (The "law of truly large numbers" is starting to come into play.) New York state has 62 counties, and it is reasonable to assume that each county has five elementary schools. Does the

event still qualify as a "statistical impossibility", or is it becoming something that could be expected to occur?

(c) If the probability in part (b) still seems small, consider further that this event could have happened in any one of the 50 states, during any of the last 10 years, and still would have received the same news coverage.

In addition to Diaconis and Mosteller (1989), see Hanley (1992) for more on coincidences.

3.10 Shuster (1991) describes a number of probability calculations that he did for a court case involving the sale of cocaine. A Florida police department seized 496 *suspected* packets of cocaine, of which four were randomly selected and tested and found to actually be cocaine. The police then chose two more packets at random and, posing as drug dealers, sold the packets to the defendant. These last two packets were lost before they could be tested to verify that they were, indeed, cocaine.

(a) If the original 496 packets were composed of N packets of cocaine and $M = 496 - N$ noncocaine, show that the probability of selecting 4 cocaine packets and then 2 noncocaine packets, which is the probability that the defendant is innocent of buying cocaine, is

$$\frac{\binom{N}{4}\binom{M}{2}}{\binom{N+M}{4}\binom{N+M-4}{2}}.$$

(b) Maximizing (in M and N) the probability in part (a) maximizes the defendant's "innocence probability". Show that this probability is .022, attained at $M = 165$ and $N = 331$.

3.11 The hypergeometric distribution can be approximated by either the binomial or the Poisson distribution. (Of course, it can be approximated by other distributions, but in this exercise we will concentrate on only these two.) Let X have the hypergeometric distribution

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

(a) Show that as $N \to \infty, M \to \infty$, and $M/N \to p$,

$$P(X = x|N, M, K) \to \binom{K}{x} p^x (1-p)^{K-x}, \quad x = 0, 1, \ldots, K.$$

(Stirling's Formula from Exercise 1.23 may be helpful.)

(b) Use the fact that the binomial can be approximated by the Poisson to show that if $N \to \infty, M \to \infty, K \to \infty$, $M/N \to 0$, and $KM/N \to \lambda$, then

$$P(X = x|N, M, K) \to \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

(c) Verify the approximation in part (b) directly, without using the Poisson approximation to the binomial. (Lemma 2.3.14 is helpful.)

3.12 Suppose X has a binomial(n,p) distribution and let Y have a negative binomial(r,p) distribution. Show that $F_X(r-1) = 1 - F_Y(n-r)$.

3.13 A *truncated* discrete distribution is one in which a particular class cannot be observed and is eliminated from the sample space. In particular, if X has range $0, 1, 2, \ldots$ and the 0 class cannot be observed (as is usually the case), the 0-*truncated* random variable X_T has pmf

$$P(X_T = x) = \frac{P(X = x)}{P(X > 0)}, \quad x = 1, 2, \ldots.$$

Find the pmf, mean, and variance of the 0-truncated random variable starting from

(a) $X \sim$ Poisson(λ).

(b) $X \sim$ negative binomial(r,p), as in (3.2.10).

3.14 Starting from the 0-truncated negative binomial (refer to Exercise 3.13), if we let $r \to 0$, we get an interesting distribution, the *logarithmic series distribution*. A random variable X has a logarithmic series distribution with parameter p if

$$P(X = x) = \frac{-(1-p)^x}{x \log p}, \quad x = 1, 2, \ldots, \quad 0 < p < 1.$$

(a) Verify that this defines a legitimate probability function.

(b) Find the mean and variance of X. (The logarithmic series distribution has proved useful in modeling species abundance. See Stuart and Ord 1987 for a more detailed discussion of this distribution.)

3.15 In Section 3.2 it was claimed that the Poisson(λ) distribution is the limit of the negative binomial(r,p) distribution as $r \to \infty$, $p \to 1$, and $r(1-p) \to \lambda$. Show that under these conditions the mgf of the negative binomial converges to that of the Poisson.

3.16 Verify these two identities regarding the gamma function that were given in the text:

(a) $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$

(b) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

3.17 Establish a formula similar to (3.3.18) for the gamma distribution. If $X \sim$ gamma(α, β), then for any positive constant ν,

$$\mathrm{E}X^\nu = \frac{\beta^\nu \Gamma(\nu + \alpha)}{\Gamma(\alpha)}.$$

3.18 There is an interesting relationship between negative binomial and gamma random variables, which may sometimes provide a useful approximation. Let Y be a negative binomial random variable with parameters r and p, where p is the success probability. Show that as $p \to 0$, the mgf of the random variable pY converges to that of a gamma distribution with parameters r and 1.

3.19 Show that

$$\int_x^\infty \frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z} dz = \sum_{y=0}^{\alpha-1} \frac{x^y e^{-x}}{y!}, \quad \alpha = 1, 2, 3, \ldots.$$

(*Hint:* Use integration by parts.) Express this formula as a probabilistic relationship between Poisson and gamma random variables.

3.20 Let the random variable X have the pdf

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad 0 < x < \infty.$$

(a) Find the mean and variance of X. (This distribution is sometimes called a *folded normal*.)

(b) If X has the folded normal distribution, find the transformation $g(X) = Y$ and values of α and β so that $Y \sim \text{gamma}(\alpha, \beta)$.

3.21 Write the integral that would define the mgf of the pdf

$$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}.$$

Is the integral finite? (Do you expect it to be?)

3.22 For each of the following distributions, verify the formulas for $\text{E}X$ and $\text{Var}\, X$ given in the text.

(a) Verify $\text{Var}\, X$ if X has a Poisson(λ) distribution. (*Hint:* Compute $\text{E}X(X-1) = \text{E}X^2 - \text{E}X$.)

(b) Verify $\text{Var}\, X$ if X has a negative binomial(r, p) distribution.

(c) Verify $\text{Var}\, X$ if X has a gamma(α, β) distribution.

(d) Verify $\text{E}X$ and $\text{Var}\, X$ if X has a beta(α, β) distribution.

(e) Verify $\text{E}X$ and $\text{Var}\, X$ if X has a double exponential(μ, σ) distribution.

3.23 The *Pareto distribution*, with parameters α and β, has pdf

$$f(x) = \frac{\beta \alpha^\beta}{x^{\beta+1}}, \quad \alpha < x < \infty, \quad \alpha > 0, \quad \beta > 0.$$

(a) Verify that $f(x)$ is a pdf.

(b) Derive the mean and variance of this distribution.

(c) Prove that the variance does not exist if $\beta \leq 2$.

3.24 Many "named" distributions are special cases of the more common distributions already discussed. For each of the following named distributions derive the form of the pdf, verify that it is a pdf, and calculate the mean and variance.

(a) If $X \sim$ exponential(β), then $Y = X^{1/\gamma}$ has the *Weibull*(γ, β) *distribution*, where $\gamma > 0$ is a constant.

(b) If $X \sim$ exponential(β), then $Y = (2X/\beta)^{1/2}$ has the *Rayleigh distribution*.

(c) If $X \sim$ gamma(a, b), then $Y = 1/X$ has the *inverted gamma* IG(a, b) distribution. (This distribution is useful in Bayesian estimation of variances; see Exercise 7.23.)

(d) If $X \sim$ gamma($\frac{3}{2}, \beta$), then $Y = (X/\beta)^{1/2}$ has the *Maxwell distribution*.

(e) If $X \sim$ exponential(1), then $Y = \alpha - \gamma \log X$ has the *Gumbel*(α, γ) *distribution*, where $-\infty < \alpha < \infty$ and $\gamma > 0$. (The Gumbel distribution is also known as the *extreme value distribution*.)

3.25 Suppose the random variable T is the length of life of an object (possibly the lifetime of an electrical component or of a subject given a particular treatment). The *hazard function* $h_T(t)$ associated with the random variable T is defined by

$$h_T(t) = \lim_{\delta \to 0} \frac{P(t \leq T < t+\delta | T \geq t)}{\delta}.$$

Thus, we can interpret $h_T(t)$ as the rate of change of the probability that the object survives a little past time t, given that the object survives to time t. Show that if T is a continuous random variable, then

$$h_T(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{dt} \log (1 - F_T(t)) .$$

3.26 Verify that the following pdfs have the indicated hazard functions (see Exercise 3.25).

(a) If $T \sim$ exponential(β), then $h_T(t) = 1/\beta$.

(b) If $T \sim$ Weibull(γ, β), then $h_T(t) = (\gamma/\beta)t^{\gamma-1}$.

(c) If $T \sim$ logistic(μ, β), that is,

$$F_T(t) = \frac{1}{1 + e^{-(t-\mu)/\beta}},$$

then $h_T(t) = (1/\beta)F_T(t)$.

3.27 For each of the following families, show whether all the pdfs in the family are unimodal (see Exercise 2.27).

(a) uniform(a, b)

(b) gamma(α, β)

(c) n(μ, σ^2)

(d) beta(α, β)

3.28 Show that each of the following families is an exponential family.

(a) normal family with either parameter μ or σ known

(b) gamma family with either parameter α or β known or both unknown

(c) beta family with either parameter α or β known or both unknown

(d) Poisson family

(e) negative binomial family with r known, $0 < p < 1$

3.29 For each family in Exercise 3.28, describe the natural parameter space.

3.30 Use the identities of Theorem 3.4.2 to

(a) calculate the variance of a binomial random variable.

(b) calculate the mean and variance of a beta(a, b) random variable.

3.31 In this exercise we will prove Theorem 3.4.2.

(a) Start from the equality

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

differentiate both sides, and then rearrange terms to establish (3.4.4). (The fact that $\frac{d}{dx} \log g(x) = g'(x)/g(x)$ will be helpful.)

(b) Differentiate the above equality a second time; then rearrange to establish (3.4.5). (The fact that $\frac{d^2}{dx^2} \log g(x) = (g''(x)/g(x)) - (g'(x)/g(x))^2$ will be helpful.)

3.32 (a) If an exponential family can be written in the form (3.4.7), show that the identities of Theorem 3.4.2 simplify to

$$\mathrm{E}(t_j(X)) = -\frac{\partial}{\partial \eta_j} \log c^*(\boldsymbol{\eta}),$$

$$\mathrm{Var}(t_j(X)) = -\frac{\partial^2}{\partial \eta_j^2} \log c^*(\boldsymbol{\eta}).$$

(b) Use this identity to calculate the mean and variance of a gamma(a, b) random variable.

3.33 For each of the following families:

(i) Verify that it is an exponential family.

(ii) Describe the curve on which the $\boldsymbol{\theta}$ parameter vector lies.

(iii) Sketch a graph of the curved parameter space.

(a) n(θ, θ)
(b) n$(\theta, a\theta^2)$, a known
(c) gamma$(\alpha, 1/\alpha)$
(d) $f(x|\theta) = C \exp\left(-(x-\theta)^4\right)$, C a normalizing constant

3.34 In Example 3.4.9 we saw that normal approximations can result in curved exponential families. For each of the following normal approximations:

(i) Describe the curve on which the $\boldsymbol{\theta}$ parameter vector lies.

(ii) Sketch a graph of the curved parameter space.

(a) Poisson approximation: $\bar{X} \sim \mathrm{n}(\lambda, \lambda/n)$
(b) binomial approximation: $\bar{X} \sim \mathrm{n}(p, p(1-p)/n)$
(c) negative binomial approximation: $\bar{X} \sim \mathrm{n}(r(1-p)/p, r(1-p)/np^2)$

3.35 (a) The normal family that approximates a Poisson can also be parameterized as n(e^θ, e^θ), where $-\infty < \theta < \infty$. Sketch a graph of the parameter space, and compare with the approximation in Exercise 3.34(a).

(b) Suppose that $X \sim$ gamma(α, β) and we assume that $\mathrm{E}X = \mu$. Sketch a graph of the parameter space.

(c) Suppose that $X_i \sim$ gamma(α_i, β_i), $i = 1, 2, \ldots, n$, and we assume that $\mathrm{E}X_i = \mu$. Describe the parameter space $(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n)$.

3.36 Consider the pdf $f(x) = \frac{63}{4}(x^6 - x^8)$, $-1 < x < 1$. Graph $(1/\sigma)f((x-\mu)/\sigma)$ for each of the following on the same axes.

(a) $\mu = 0$, $\sigma = 1$
(b) $\mu = 3$, $\sigma = 1$
(c) $\mu = 3$, $\sigma = 2$

3.37 Show that if $f(x)$ is a pdf, symmetric about 0, then μ is the median of the location–scale pdf $(1/\sigma)f((x-\mu)/\sigma)$, $-\infty < x < \infty$.

3.38 Let Z be a random variable with pdf $f(z)$. Define z_α to be a number that satisfies this relationship:

$$\alpha = P(Z > z_\alpha) = \int_{z_\alpha}^{\infty} f(z)dz.$$

Show that if X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ and $x_\alpha = \sigma z_\alpha + \mu$, then $P(X > x_\alpha) = \alpha$. (Thus if a table of z_α values were available, then values of x_α could be easily computed for any member of the location–scale family.)

3.39 Consider the Cauchy family defined in Section 3.3. This family can be extended to a location–scale family yielding pdfs of the form

$$f(x|\mu,\sigma) = \frac{1}{\sigma\pi\left(1+\left(\frac{x-\mu}{\sigma}\right)^2\right)}, \quad -\infty < x < \infty.$$

The mean and variance do not exist for the Cauchy distribution. So the parameters μ and σ^2 are not the mean and variance. But they do have important meaning. Show that if X is a random variable with a Cauchy distribution with parameters μ and σ, then:

(a) μ is the median of the distribution of X, that is, $P(X \geq \mu) = P(X \leq \mu) = \frac{1}{2}$.

(b) $\mu+\sigma$ and $\mu-\sigma$ are the quartiles of the distribution of X, that is, $P(X \geq \mu+\sigma) = P(X \leq \mu-\sigma) = \frac{1}{4}$. (*Hint:* Prove this first for $\mu = 0$ and $\sigma = 1$ and then use Exercise 3.38.)

3.40 Let $f(x)$ be any pdf with mean μ and variance σ^2. Show how to create a location–scale family based on $f(x)$ such that the standard pdf of the family, say $f^*(x)$, has mean 0 and variance 1.

3.41 A family of cdfs $\{F(x|\theta), \theta \in \Theta\}$ is *stochastically increasing in* θ if $\theta_1 > \theta_2 \Rightarrow F(x|\theta_1)$ is stochastically greater than $F(x|\theta_2)$. (See Exercise 1.49 for the definition of stochastically greater.)

(a) Show that the $n(\mu, \sigma^2)$ family is stochastically increasing in μ for fixed σ^2.

(b) Show that the gamma(α, β) family of (3.3.6) is stochastically increasing in β (scale parameter) for fixed α (shape parameter).

3.42 Refer to Exercise 3.41 for the definition of a stochastically increasing family.

(a) Show that a location family is stochastically increasing in its location parameter.

(b) Show that a scale family is stochastically increasing in its scale parameter if the sample space is $[0, \infty)$.

3.43 A family of cdfs $\{F(x|\theta), \theta \in \theta\}$ is *stochastically decreasing in* θ if $\theta_1 > \theta_2 \Rightarrow F(x|\theta_2)$ is stochastically greater than $F(x|\theta_1)$. (See Exercises 3.41 and 3.42.)

(a) Prove that if $X \sim F_X(x|\theta)$, where the sample space of X is $(0, \infty)$ and $F_X(x|\theta)$ is stochastically increasing in θ, then $F_Y(y|\theta)$ is stochastically decreasing in θ, where $Y = 1/X$.

(b) Prove that if $X \sim F_X(x|\theta)$, where $F_X(x|\theta)$ is stochastically increasing in θ and $\theta > 0$, then $F_X(x|\frac{1}{\theta})$ is stochastically decreasing in θ.

3.44 For any random variable X for which EX^2 and $E|X|$ exist, show that $P(|X| \geq b)$ does not exceed either EX^2/b^2 or $E|X|/b$, where b is a positive constant. If $f(x) = e^{-x}$ for $x > 0$, show that one bound is better when $b = 3$ and the other when $b = \sqrt{2}$. (Notice Markov's Inequality in Miscellanea 3.8.2.)

3.45 Let X be a random variable with moment-generating function $M_X(t)$, $-h < t < h$.

(a) Prove that $P(X \geq a) \leq e^{-at}M_X(t)$, $0 < t < h$. (A proof similar to that used for Chebychev's Inequality will work.)

(b) Similarly, prove that $P(X \le a) \le e^{-at} M_X(t)$, $\quad -h < t < 0$.

(c) A special case of part (a) is that $P(X \ge 0) \le \mathrm{E}e^{tX}$ for all $t \ge 0$ for which the mgf is defined. What are general conditions on a function $h(t,x)$ such that $P(X \ge 0) \le \mathrm{E}h(t,X)$ for all $t \ge 0$ for which $\mathrm{E}h(t,X)$ exists? (In part (a), $h(t,x) = e^{tx}$.)

3.46 Calculate $P(|X - \mu_X| \ge k\sigma_X)$ for $X \sim$ uniform$(0,1)$ and $X \sim$ exponential(λ), and compare your answers to the bound from Chebychev's Inequality.

3.47 If Z is a standard normal random variable, prove this companion to the inequality in Example 3.6.3:

$$P(|Z| \ge t) \ge \sqrt{\frac{2}{\pi}} \frac{t}{1+t^2} e^{-t^2/2}.$$

3.48 Derive recursion relations, similar to the one given in (3.6.2), for the binomial, negative binomial, and hypergeometric distributions.

3.49 Prove the following analogs to Stein's Lemma, assuming appropriate conditions on the function g.

(a) If $X \sim$ gamma(α, β), then

$$\mathrm{E}\left(g(X)(X - \alpha\beta)\right) = \beta \mathrm{E}\left(Xg'(X)\right).$$

(b) If $X \sim$ beta(α, β), then

$$\mathrm{E}\left[g(X)\left(\beta - (\alpha - 1)\frac{(1-X)}{X}\right)\right] = \mathrm{E}\left((1-X)g'(X)\right).$$

3.50 Prove the identity for the negative binomial distribution given in Theorem 3.6.8, part (b).

3.8 Miscellanea

3.8.1 The Poisson Postulates

The Poisson distribution can be derived from a set of basic assumptions, sometimes called the Poisson postulates. These assumptions relate to the physical properties of the process under consideration. While, generally speaking, the assumptions are not very easy to verify, they do provide an experimenter with a set of guidelines for considering whether the Poisson will provide a reasonable model. For a more complete treatment of the Poisson postulates, see the classic text by Feller (1968) or Barr and Zehna (1983).

Theorem 3.8.1 *For each $t \ge 0$, let N_t be an integer-valued random variable with the following properties. (Think of N_t as denoting the number of arrivals in the time period from time 0 to time t.)*

i) $N_0 = 0$ *(start with no arrivals)*

ii) $s < t \Rightarrow N_s$ and $N_t - N_s$ *are independent.* *(arrivals in disjoint time periods are independent)*

iii) N_s and $N_{t+s} - N_t$ *are identically distributed.* (*number of arrivals depends only on period length*)

iv) $\lim_{t\to 0} \frac{P(N_t = 1)}{t} = \lambda$ (*arrival probability proportional to period length, if length is small*)

v) $\lim_{t\to 0} \frac{P(N_t > 1)}{t} = 0$ (*no simultaneous arrivals*)

If i–v hold, then for any integer n,

$$P(N_t = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!},$$

that is, $N_t \sim$ *Poisson*(λt).

The postulates may also be interpreted as describing the behavior of objects spatially (for example, movement of insects), giving the Poisson application in spatial distributions.

3.8.2 Chebychev and Beyond

Ghosh and Meeden (1977) discuss the fact that Chebychev's Inequality is very conservative and is almost never attained. If we write $\bar{X}_n$ for the mean of the random variables $X_1, X_2, \ldots, X_n$, then Chebychev's Inequality states

$$P\left(|\bar{X}_n - \mu| \geq k\sigma\right) \leq \frac{1}{nk^2}.$$

They prove the following theorem.

Theorem 3.8.2 *If* $0 < \sigma < \infty$, *then*

a. *If* $n = 1$, *the inequality is attainable for* $k \geq 1$ *and unattainable for* $0 < k < 1$.
b. *If* $n = 2$, *the inequality is attainable if and only if* $k = 1$.
c. *If* $n \geq 3$, *the inequality is not attainable.*

Examples are given for the cases when the inequality is attained. Most of their technical arguments are based on the following inequality, known as Markov's Inequality.

Lemma 3.8.3 (Markov's Inequality) *If* $P(Y \geq 0) = 1$ *and* $P(Y = 0) < 1$, *then, for any* $r > 0$,

$$P(Y \geq r) \leq \frac{\mathrm{E}Y}{r}$$

with equality if and only if $P(Y = r) = p = 1 - P(Y = 0), 0 < p \leq 1$.

Markov's Inequality can then be applied to the quantity

$$Y = \frac{(\bar{X}_n - \mu)^2}{\sigma^2}$$

to get the above results.

One reason Chebychev's Inequality is so loose is that it puts no restrictions on the underlying distribution. With the additional restriction of *unimodality*, we can get tighter bounds and the inequalities of Gauss and Vysochanskiĭ-Petunin. (See Pukelsheim 1994 for details and elementary calculus-based proofs of these inequalities.)

Theorem 3.8.4 (Gauss Inequality) *Let $X \sim f$, where f is unimodal with mode ν, and define $\tau^2 = \mathrm{E}(X-\nu)^2$. Then*

$$P(|X-\nu| > \varepsilon) \le \begin{cases} \frac{4\tau^2}{9\varepsilon^2} & \text{for all } \varepsilon \ge \sqrt{4/3}\tau \\ 1 - \frac{\varepsilon}{\sqrt{3}\tau} & \text{for all } \varepsilon \le \sqrt{4/3}\tau. \end{cases}$$

Although this is a tighter bound than Chebychev, the dependence on the mode limits its usefulness. The extension of Vysochanskiĭ-Petunin removes this limitation.

Theorem 3.8.5 (Vysochanskiĭ-Petunin Inequality) *Let $X \sim f$, where f is unimodal, and define $\xi^2 = \mathrm{E}(X-\alpha)^2$ for an arbitrary point α. Then*

$$P(|X-\alpha| > \varepsilon) \le \begin{cases} \frac{4\xi^2}{9\varepsilon^2} & \text{for all } \varepsilon \ge \sqrt{8/3}\xi \\ \frac{4\xi^2}{9\varepsilon^2} - \frac{1}{3} & \text{for all } \varepsilon \le \sqrt{8/3}\xi. \end{cases}$$

Pukelsheim points out that taking $\alpha = \mu = \mathrm{E}(X)$ and $\varepsilon = 3\sigma$, where $\sigma^2 = \mathrm{Var}(X)$, yields

$$P(|X-\mu| > 3\sigma) \le \frac{4}{81} < .05,$$

the so-called *three-sigma rule*, that the probability is less than 5% that X is more than three standard deviations from the mean of the population.

3.8.3 More on Exponential Families

Is the lognormal distribution in the exponential family? The density given in (3.3.21) can be put into the form specified by (3.4.1). Hence, we have put the lognormal into the exponential family.

According to Brown (1986, Section 1.1), to define an exponential family of distributions we start with a nonnegative function $\nu(x)$ and define the set $\mathcal{N}$ by

$$\mathcal{N} = \left\{ \theta : \int_{\mathcal{X}} e^{\theta x} \nu(x)\, dx < \infty \right\}.$$

If we let $\lambda(\theta) = \int_{\mathcal{X}} e^{\theta x}\nu(x)\, dx$, the set of probability densities defined by

$$f(x|\theta) = \frac{e^{\theta x}\nu(x)}{\lambda(\theta)}, \quad x \in \mathcal{X}, \quad \theta \in \mathcal{N},$$

is an *exponential family*. The moment-generating function of $f(x|\theta)$ is

$$M_X(t) = \int_{\mathcal{X}} e^{tx} f(x|\theta)\, dx = \frac{\lambda(t+\theta)}{\lambda(\theta)}$$

and hence exists by construction. If the parameter space Θ is equal to the set $\mathcal{N}$, the exponential family is called *full*. Cases where Θ is a lower-dimensional subset of $\mathcal{N}$ give rise to curved exponential families.

Returning to the lognormal distribution, we know that it does not have an mgf, so it can't satisfy Brown's definition of an exponential family. However, the lognormal satisfies the expectation identities of Theorem 3.4.2 and enjoys the sufficiency properties detailed in Section 6.2.1 (Theorem 6.2.10). For our purposes, these are the major properties that we need and the main reasons for identifying a member of the exponential family. More advanced properties, which we will not investigate here, may need the existence of the mgf.