Chapter 2

Transformations and Expectations

"We want something more than mere theory and preaching now, though."
Sherlock Holmes
A Study in Scarlet

Often, if we are able to model a phenomenon in terms of a random variable X with cdf $F_X(x)$, we will also be concerned with the behavior of functions of X. In this chapter we study techniques that allow us to gain information about functions of X that may be of interest, information that can range from very complete (the distributions of these functions) to more vague (the average behavior).

2.1 Distributions of Functions of a Random Variable

If X is a random variable with cdf $F_X(x)$, then any function of X, say $g(X)$, is also a random variable. Often $g(X)$ is of interest itself and we write $Y = g(X)$ to denote the new random variable $g(X)$. Since Y is a function of X, we can describe the probabilistic behavior of Y in terms of that of X. That is, for any set A,

$$P(Y \in A) = P(g(X) \in A),$$

showing that the distribution of Y depends on the functions F_X and g. Depending on the choice of g, it is sometimes possible to obtain a tractable expression for this probability.

Formally, if we write $y = g(x)$, the function $g(x)$ defines a mapping from the original sample space of X, $\mathcal{X}$, to a new sample space, $\mathcal{Y}$, the sample space of the random variable Y. That is,

$$g(x)\colon\ \mathcal{X} \to \mathcal{Y}.$$

We associate with g an inverse mapping, denoted by g^{-1}, which is a mapping from subsets of $\mathcal{Y}$ to subsets of $\mathcal{X}$, and is defined by

$$g^{-1}(A) = \{x \in \mathcal{X}\colon\ g(x) \in A\}. \tag{2.1.1}$$

Note that the mapping g^{-1} takes sets into sets; that is, $g^{-1}(A)$ is the set of points in $\mathcal{X}$ that $g(x)$ takes into the set A. It is possible for A to be a point set, say $A = \{y\}$. Then

$$g^{-1}(\{y\}) = \{x \in \mathcal{X}\colon\ g(x) = y\}.$$

In this case we often write $g^{-1}(y)$ instead of $g^{-1}(\{y\})$. The quantity $g^{-1}(y)$ can still be a set, however, if there is more than one x for which $g(x) = y$. If there is only one x for which $g(x) = y$, then $g^{-1}(y)$ is the point set $\{x\}$, and we will write $g^{-1}(y) = x$. If the random variable Y is now defined by $Y = g(X)$, we can write for any set $A \subset \mathcal{Y}$,

$$\begin{aligned} P(Y \in A) &= P(g(X) \in A) \\ &= P(\{x \in \mathcal{X}\colon\ g(x) \in A\}) \\ &= P\left(X \in g^{-1}(A)\right). \end{aligned} \tag{2.1.2}$$

This defines the probability distribution of Y. It is straightforward to show that this probability distribution satisfies the Kolmogorov Axioms.

If X is a discrete random variable, then $\mathcal{X}$ is countable. The sample space for $Y = g(X)$ is $\mathcal{Y} = \{y\colon\ y = g(x),\ x \in \mathcal{X}\}$, which is also a countable set. Thus, Y is also a discrete random variable. From (2.1.2), the pmf for Y is

$$f_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x), \quad \text{for } y \in \mathcal{Y},$$

and $f_Y(y) = 0$ for $y \notin \mathcal{Y}$. In this case, finding the pmf of Y involves simply identifying $g^{-1}(y)$, for each $y \in \mathcal{Y}$, and summing the appropriate probabilities.

Example 2.1.1 (Binomial transformation) A discrete random variable X has a *binomial distribution* if its pmf is of the form

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \tag{2.1.3}$$

where n is a positive integer and $0 \le p \le 1$. Values such as n and p that can be set to different values, producing different probability distributions, are called *parameters*. Consider the random variable $Y = g(X)$, where $g(x) = n - x$; that is, $Y = n - X$. Here $\mathcal{X} = \{0, 1, \ldots, n\}$ and $\mathcal{Y} = \{y\colon y = g(x),\ x \in \mathcal{X}\} = \{0, 1, \ldots, n\}$. For any $y \in \mathcal{Y}$, $n - x = g(x) = y$ if and only if $x = n - y$. Thus, $g^{-1}(y)$ is the single point $x = n - y$, and

$$\begin{aligned} f_Y(y) &= \sum_{x \in g^{-1}(y)} f_X(x) \\ &= f_X(n - y) \\ &= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} \\ &= \binom{n}{y} (1-p)^y p^{n-y}. \qquad \left(\begin{array}{c}\text{Definition 1.2.17} \\ \text{implies } \binom{n}{y} = \binom{n}{n-y}\end{array}\right) \end{aligned}$$

Thus, we see that Y also has a binomial distribution, but with parameters n and $1 - p$. ‖

If X and Y are continuous random variables, then in some cases it is possible to find simple formulas for the cdf and pdf of Y in terms of the cdf and pdf of X and the function g. In the remainder of this section, we consider some of these cases.

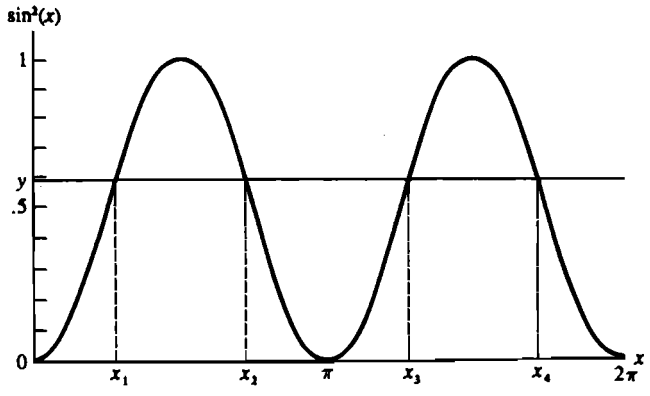


Figure 2.1.1. *Graph of the transformation* $y = \sin^2(x)$ *of Example 2.1.2*

The cdf of $Y = g(X)$ is

$$
\begin{aligned}
F_Y(y) &= P(Y \le y) \\
&= P(g(X) \le y) \\
&= P(\{x \in \mathcal{X}\colon\ g(x) \le y\}) \\
&= \int_{\{x \in \mathcal{X}\colon\ g(x) \le y\}} f_X(x)\,dx.
\end{aligned} \tag{2.1.4}
$$

Sometimes there may be difficulty in identifying $\{x \in \mathcal{X}\colon\ g(x) \le y\}$ and carrying out the integration of $f_X(x)$ over this region, as the next example shows.

Example 2.1.2 (Uniform transformation) Suppose X has a uniform distribution on the interval $(0, 2\pi)$, that is,

$$
f_X(x) = \begin{cases} 1/(2\pi) & 0 < x < 2\pi \\ 0 & \text{otherwise.} \end{cases}
$$

Consider $Y = \sin^2(X)$. Then (see Figure 2.1.1)

$$
P(Y \le y) = P(X \le x_1) + P(x_2 \le X \le x_3) + P(X \ge x_4). \tag{2.1.5}
$$

From the symmetry of the function $\sin^2(x)$ and the fact that X has a uniform distribution, we have

$$
P(X \le x_1) = P(X \ge x_4) \quad \text{and} \quad P(x_2 \le X \le x_3) = 2P(x_2 \le X \le \pi),
$$

so

$$
P(Y \le y) = 2P(X \le x_1) + 2P(x_2 \le X \le \pi), \tag{2.1.6}
$$

where x_1 and x_2 are the two solutions to

$$
\sin^2(x) = y, \qquad 0 < x < \pi.
$$

Thus, even though this example dealt with a seemingly simple situation, the resulting expression for the cdf of Y was not simple. ||

When transformations are made, it is important to keep track of the sample spaces of the random variables; otherwise, much confusion can arise. When the transformation is from X to $Y = g(X)$, it is most convenient to use

$$\text{(2.1.7)} \quad \mathcal{X} = \{x\colon\ f_X(x) > 0\} \qquad \text{and} \qquad \mathcal{Y} = \{y\colon\ y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

The pdf of the random variable X is positive only on the set $\mathcal{X}$ and is 0 elsewhere. Such a set is called the *support set* of a distribution or, more informally, the *support* of a distribution. This terminology can also apply to a pmf or, in general, to any nonnegative function.

It is easiest to deal with functions $g(x)$ that are *monotone*, that is, those that satisfy either

$$u > v \Rightarrow g(u) > g(v) \quad \text{(increasing)} \quad \text{or} \quad u < v \Rightarrow g(u) > g(v) \quad \text{(decreasing)}.$$

If the transformation $x \to g(x)$ is monotone, then it is *one-to-one and onto* from $\mathcal{X} \to \mathcal{Y}$. That is, each x goes to only one y and each y comes from at most one x (one-to-one). Also, for $\mathcal{Y}$ defined as in (2.1.7), for each $y \in \mathcal{Y}$ there is an $x \in \mathcal{X}$ such that $g(x) = y$ (onto). Thus, the transformation g uniquely pairs xs and ys. If g is monotone, then g^{-1} is single-valued; that is, $g^{-1}(y) = x$ if and only if $y = g(x)$. If g is increasing, this implies that

$$\begin{aligned} \{x \in \mathcal{X}\colon\ g(x) \le y\} &= \{x \in \mathcal{X}\colon\ g^{-1}(g(x)) \le g^{-1}(y)\} \\ &= \{x \in \mathcal{X}\colon\ x \le g^{-1}(y)\}. \end{aligned} \tag{2.1.8}$$

If g is decreasing, this implies that

$$\begin{aligned} \{x \in \mathcal{X}\colon\ g(x) \le y\} &= \{x \in \mathcal{X}\colon\ g^{-1}(g(x)) \ge g^{-1}(y)\} \\ &= \{x \in \mathcal{X}\colon\ x \ge g^{-1}(y)\}. \end{aligned} \tag{2.1.9}$$

(A graph will illustrate why the inequality reverses in the decreasing case.) If $g(x)$ is an increasing function, then using (2.1.4), we can write

$$F_Y(y) = \int_{\{x \in \mathcal{X}\colon\ x \le g^{-1}(y)\}} f_X(x)\,dx = \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx = F_X\left(g^{-1}(y)\right).$$

If $g(x)$ is decreasing, we have

$$F_Y(y) = \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx = 1 - F_X\left(g^{-1}(y)\right).$$

The continuity of X is used to obtain the second equality. We summarize these results in the following theorem.

Theorem 2.1.3 *Let X have cdf $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as in (2.1.7).*

a. *If g is an increasing function on $\mathcal{X}$, $F_Y(y) = F_X\left(g^{-1}(y)\right)$ for $y \in \mathcal{Y}$.*

b. *If g is a decreasing function on $\mathcal{X}$ and X is a continuous random variable, $F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$ for $y \in \mathcal{Y}$.*

Example 2.1.4 (Uniform-exponential relationship–I) Suppose $X \sim f_X(x) = 1$ if $0 < x < 1$ and 0 otherwise, the *uniform*$(0,1)$ *distribution*. It is straightforward to check that $F_X(x) = x$, $0 < x < 1$. We now make the transformation $Y = g(X) = -\log X$. Since

$$\frac{d}{dx}g(x) = \frac{d}{dx}(-\log x) = \frac{-1}{x} < 0, \qquad \text{for} \quad 0 < x < 1,$$

$g(x)$ is a decreasing function. As X ranges between 0 and 1, $-\log x$ ranges between 0 and ∞, that is, $\mathcal{Y} = (0, \infty)$. For $y > 0$, $y = -\log x$ implies $x = e^{-y}$, so $g^{-1}(y) = e^{-y}$. Therefore, for $y > 0$,

$$F_Y(y) = 1 - F_X\left(g^{-1}(y)\right) = 1 - F_X(e^{-y}) = 1 - e^{-y}. \qquad (F_X(x) = x)$$

Of course, $F_Y(y) = 0$ for $y \leq 0$. Note that it was necessary only to verify that $g(x) = -\log x$ is monotone on $(0,1)$, the support of X. ‖

If the pdf of Y is continuous, it can be obtained by differentiating the cdf. The resulting expression is given in the following theorem.

Theorem 2.1.5 *Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X}$ and $\mathcal{Y}$ be defined by (2.1.7). Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y))\left|\dfrac{d}{dy}g^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \textit{otherwise.} \end{cases} \tag{2.1.10}$$

Proof: From Theorem 2.1.3 we have, by the chain rule,

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \begin{cases} f_X(g^{-1}(y))\dfrac{d}{dy}g^{-1}(y) & \text{if } g \text{ is increasing} \\ -f_X(g^{-1}(y))\dfrac{d}{dy}g^{-1}(y) & \text{if } g \text{ is decreasing,} \end{cases}$$

which can be expressed concisely as (2.1.10). □

Example 2.1.6 (Inverted gamma pdf) Let $f_X(x)$ be the *gamma pdf*

$$f(x) = \frac{1}{(n-1)!\beta^n}x^{n-1}e^{-x/\beta}, \qquad 0 < x < \infty,$$

where β is a positive constant and n is a positive integer. Suppose we want to find the pdf of $g(X) = 1/X$. Note that here the support sets $\mathcal{X}$ and $\mathcal{Y}$ are both the interval

$(0, \infty)$. If we let $y = g(x)$, then $g^{-1}(y) = 1/y$ and $\frac{d}{dy}g^{-1}(y) = -1/y^2$. Applying the above theorem, for $y \in (0, \infty)$, we get

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right| \\ &= \frac{1}{(n-1)!\beta^n}\left(\frac{1}{y}\right)^{n-1} e^{-1/(\beta y)}\frac{1}{y^2} \\ &= \frac{1}{(n-1)!\beta^n}\left(\frac{1}{y}\right)^{n+1} e^{-1/(\beta y)}, \end{aligned}$$

a special case of a pdf known as the *inverted gamma pdf.* ||

In many applications, the function g may be neither increasing nor decreasing; hence the above results will not apply. However, it is often the case that g will be monotone over certain intervals and that allows us to get an expression for $Y = g(X)$. (If g is not monotone over certain intervals, then we are in *deep* trouble.)

Example 2.1.7 (Square transformation) Suppose X is a continuous random variable. For $y > 0$, the cdf of $Y = X^2$ is

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y}).$$

Because x is continuous, we can drop the equality from the left endpoint and obtain

$$\begin{aligned} F_Y(y) &= P(-\sqrt{y} < X \le \sqrt{y}) \\ &= P(X \le \sqrt{y}) - P(X \le -\sqrt{y}) \quad = \quad F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

The pdf of Y can now be obtained from the cdf by differentiation:

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) \\ &= \frac{d}{dy}[F_X(\sqrt{y}) - F_X(-\sqrt{y})] \\ &= \frac{1}{2\sqrt{y}}f_X(\sqrt{y}) + \frac{1}{2\sqrt{y}}f_X(-\sqrt{y}), \end{aligned}$$

where we use the chain rule to differentiate $F_X(\sqrt{y})$ and $F_X(-\sqrt{y})$. Therefore, the pdf is

$$f_Y(y) = \frac{1}{2\sqrt{y}}\left(f_X(\sqrt{y}) + f_X(-\sqrt{y})\right). \tag{2.1.11}$$

Notice that the pdf of Y in (2.1.11) is expressed as the sum of two pieces, pieces that represent the intervals where $g(x) = x^2$ is monotone. In general, this will be the case. ||

Theorem 2.1.8 *Let X have pdf $f_X(x)$, let $Y = g(X)$, and define the sample space $\mathcal{X}$ as in (2.1.7). Suppose there exists a partition, $A_0, A_1, \ldots, A_k$, of $\mathcal{X}$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i. Further, suppose there exist functions $g_1(x), \ldots, g_k(x)$, defined on $A_1, \ldots, A_k$, respectively, satisfying*

i. *$g(x) = g_i(x)$, for $x \in A_i$,*

ii. *$g_i(x)$ is monotone on A_i,*

iii. *the set $\mathcal{Y} = \{y\colon\ y = g_i(x)$ for some $x \in A_i\}$ is the same for each $i = 1, \ldots, k$, and*

iv. *$g_i^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, for each $i = 1, \ldots, k$.*

Then

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X\left(g_i^{-1}(y)\right) \left|\dfrac{d}{dy} g_i^{-1}(y)\right| & y \in \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases}$$

The important point in Theorem 2.1.8 is that $\mathcal{X}$ can be divided into sets $A_1, \ldots, A_k$ such that $g(x)$ is monotone on each A_i. We can ignore the "exceptional set" A_0 since $P(X \in A_0) = 0$. It is a technical device that is used, for example, to handle endpoints of intervals. It is important to note that each $g_i(x)$ is a one-to-one transformation from A_i onto $\mathcal{Y}$. Furthermore, $g_i^{-1}(y)$ is a one-to-one function from $\mathcal{Y}$ onto A_i such that, for $y \in \mathcal{Y}$, $g_i^{-1}(y)$ gives the unique $x = g_i^{-1}(y) \in A_i$ for which $g_i(x) = y$. (See Exercise 2.7 for an extension.)

Example 2.1.9 (Normal-chi squared relationship) Let X have the *standard normal distribution*,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

Consider $Y = X^2$. The function $g(x) = x^2$ is monotone on $(-\infty, 0)$ and on $(0, \infty)$. The set $\mathcal{Y} = (0, \infty)$. Applying Theorem 2.1.8, we take

$$\begin{aligned} &A_0 = \{0\}; \\ &A_1 = (-\infty, 0), \qquad g_1(x) = x^2, \qquad g_1^{-1}(y) = -\sqrt{y}; \\ &A_2 = (0, \infty), \qquad g_2(x) = x^2, \qquad g_2^{-1}(y) = \sqrt{y}. \end{aligned}$$

The pdf of Y is

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \left|-\frac{1}{2\sqrt{y}}\right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \left|\frac{1}{2\sqrt{y}}\right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}, \qquad 0 < y < \infty. \end{aligned}$$

The pdf of Y is one that we will often encounter, that of a *chi squared random variable* with 1 degree of freedom. ||

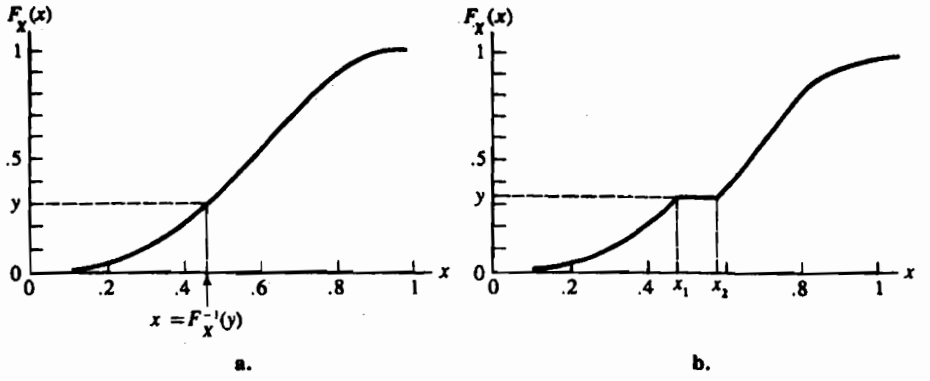


Figure 2.1.2. **(a)** $F(x)$ *strictly increasing;* **(b)** $F(x)$ *nondecreasing*

We close this section with a special and very useful transformation.

Theorem 2.1.10 (Probability integral transformation) *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0,1)$, that is, $P(Y \leq y) = y, 0 < y < 1$.*

Before we prove this theorem, we will digress for a moment and look at F_X^{-1}, the inverse of the cdf F_X, in some detail. If F_X is strictly increasing, then F_X^{-1} is well defined by

$$F_X^{-1}(y) = x \quad \Leftrightarrow \quad F_X(x) = y. \tag{2.1.12}$$

However, if F_X is constant on some interval, then F_X^{-1} is not well defined by (2.1.12), as Figure 2.1.2 illustrates. Any x satisfying $x_1 \leq x \leq x_2$ satisfies $F_X(x) = y$.

This problem is avoided by defining $F_X^{-1}(y)$ for $0 < y < 1$ by

$$F_X^{-1}(y) = \inf\{x:\ F_X(x) \geq y\}, \tag{2.1.13}$$

a definition that agrees with (2.1.12) when F_X is nonconstant and provides an F_X^{-1} that is single-valued even when F_X is not strictly increasing. Using this definition, in Figure 2.1.2b, we have $F_X^{-1}(y) = x_1$. At the endpoints of the range of y, $F_X^{-1}(y)$ can also be defined. $F_X^{-1}(1) = \infty$ if $F_X(x) < 1$ for all x and, for any F_X, $F_X^{-1}(0) = -\infty$.

Proof of Theorem 2.1.10: For $Y = F_X(X)$ we have, for $0 < y < 1$,

$$\begin{aligned}
P(Y \leq y) &= P(F_X(X) \leq y) \\
&= P(F_X^{-1}[F_X(X)] \leq F_X^{-1}(y)) && (F_X^{-1} \text{ is increasing}) \\
&= P(X \leq F_X^{-1}(y)) && (\text{see paragraph below}) \\
&= F_X\left(F_X^{-1}(y)\right) && (\text{definition of } F_X) \\
&= y. && (\text{continuity of } F_X)
\end{aligned}$$

At the endpoints we have $P(Y \le y) = 1$ for $y \ge 1$ and $P(Y \le y) = 0$ for $y \le 0$, showing that Y has a uniform distribution.

The reasoning behind the equality

$$P\left(F_X^{-1}(F_X(X)) \le F_X^{-1}(y)\right) = P(X \le F_X^{-1}(y))$$

is somewhat subtle and deserves additional attention. If F_X is strictly increasing, then it is true that $F_X^{-1}(F_X(x)) = x$. (Refer to Figure 2.1.2a.) However, if F_X is flat, it may be that $F_X^{-1}(F_X(x)) \neq x$. Suppose F_X is as in Figure 2.1.2b and let $x \in [x_1, x_2]$. Then $F_X^{-1}(F_X(x)) = x_1$ for any x in this interval. Even in this case, though, the probability equality holds, since $P(X \le x) = P(X \le x_1)$ for any $x \in [x_1, x_2]$. The flat cdf denotes a region of 0 probability ($P(x_1 < X \le x) = F_X(x) - F_X(x_1) = 0$). □

One application of Theorem 2.1.10 is in the generation of random samples from a particular distribution. If it is required to generate an observation X from a population with cdf F_X, we need only generate a uniform random number V, between 0 and 1, and solve for x in the equation $F_X(x) = u$. (For many distributions there are other methods of generating observations that take less computer time, but this method is still useful because of its general applicability.)

2.2 Expected Values

The expected value, or expectation, of a random variable is merely its average value, where we speak of "average" value as one that is weighted according to the probability distribution. The expected value of a distribution can be thought of as a measure of center, as we think of averages as being middle values. By weighting the values of the random variable according to the probability distribution, we hope to obtain a number that summarizes a typical or expected value of an observation of the random variable.

Definition 2.2.1 The *expected value* or *mean* of a random variable $g(X)$, denoted by $\mathrm{E}\,g(X)$, is

$$\mathrm{E}\,g(X) = \begin{cases} \int_{-\infty}^{\infty} g(x) f_X(x)\,dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) P(X = x) & \text{if } X \text{ is discrete,} \end{cases}$$

provided that the integral or sum exists. If $\mathrm{E}|g(X)| = \infty$, we say that $\mathrm{E}\,g(X)$ does not exist. (Ross 1988 refers to this as the "law of the unconscious statistician." We do not find this amusing.)

Example 2.2.2 (Exponential mean) Suppose X has an *exponential* (λ) *distribution*, that is, it has pdf given by

$$f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}, \qquad 0 \le x < \infty, \quad \lambda > 0.$$

Then $\mathrm{E}\,X$ is given by

$$\begin{aligned}\mathrm{E}\,X &= \int_0^\infty \frac{1}{\lambda} x e^{-x/\lambda}\,dx \\ &= -xe^{-x/\lambda}\Big|_0^\infty + \int_0^\infty e^{-x/\lambda}\,dx \qquad \text{(integration by parts)} \\ &= \int_0^\infty e^{-x/\lambda}\,dx \;=\; \lambda. \end{aligned}$$

‖

Example 2.2.3 (Binomial mean) If X has a *binomial distribution*, its pmf is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n,$$

where n is a positive integer, $0 \le p \le 1$, and for every fixed pair n and p the pmf sums to 1. The expected value of a binomial random variable is given by

$$\mathrm{E}\,X = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

(the $x = 0$ term is 0). Using the identity $x\binom{n}{x} = n\binom{n-1}{x-1}$, we have

$$\begin{aligned}\mathrm{E}\,X &= \sum_{x=1}^{n} n \binom{n-1}{x-1} p^x (1-p)^{n-x} \\ &= \sum_{y=0}^{n-1} n \binom{n-1}{y} p^{y+1} (1-p)^{n-(y+1)} \qquad \text{(substitute } y = x-1\text{)} \\ &= np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y} \\ &= np, \end{aligned}$$

since the last summation must be 1, being the sum over all possible values of a binomial$(n-1,\ p)$ pmf. ‖

Example 2.2.4 (Cauchy mean) A classic example of a random variable whose expected value does not exist is a *Cauchy random variable*, that is, one with pdf

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \qquad -\infty < x < \infty.$$

It is straightforward to check that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$, but $\mathrm{E}|X| = \infty$. Write

$$\mathrm{E}|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi} \frac{1}{1+x^2}\,dx = \frac{2}{\pi} \int_0^\infty \frac{x}{1+x^2}\,dx.$$

For any positive number M,

$$\int_0^M \frac{x}{1+x^2}\,dx = \frac{\log(1+x^2)}{2}\bigg|_0^M = \frac{\log(1+M^2)}{2}.$$

Thus,

$$\mathrm{E}|X| = \lim_{M\to\infty} \frac{2}{\pi}\int_0^M \frac{x}{1+x^2}\,dx = \frac{1}{\pi}\lim_{M\to\infty}\log(1+M^2) = \infty$$

and $\mathrm{E}\,X$ does not exist. ||

The process of taking expectations is a linear operation, which means that the expectation of a linear function of X can be easily evaluated by noting that for any constants a and b,

$$\mathrm{E}(aX+b) = a\mathrm{E}\,X + b. \tag{2.2.1}$$

For example, if X is binomial(n,p), so that $\mathrm{E}\,X = np$, then

$$\mathrm{E}(X-np) = \mathrm{E}\,X - np = np - np = 0.$$

The expectation operator, in fact, has many properties that can help ease calculational effort. Most of these properties follow from the properties of the integral or sum, and are summarized in the following theorem.

Theorem 2.2.5 *Let X be a random variable and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

a. $\mathrm{E}(ag_1(X)+bg_2(X)+c) = a\mathrm{E}\,g_1(X) + b\mathrm{E}\,g_2(X) + c.$

b. *If $g_1(x) \geq 0$ for all x, then $\mathrm{E}\,g_1(X) \geq 0$.*

c. *If $g_1(x) \geq g_2(x)$ for all x, then $\mathrm{E}\,g_1(X) \geq \mathrm{E}\,g_2(X)$.*

d. *If $a \leq g_1(x) \leq b$ for all x, then $a \leq \mathrm{E}\,g_1(X) \leq b$.*

Proof: We will give details for only the continuous case, the discrete case being similar. By definition,

$$\begin{aligned}
&\mathrm{E}(ag_1(X)+bg_2(X)+c)\\
&\quad = \int_{-\infty}^{\infty}(ag_1(x)+bg_2(x)+c)f_X(x)\,dx\\
&\quad = \int_{-\infty}^{\infty} ag_1(x)f_X(x)\,dx + \int_{-\infty}^{\infty} bg_2(x)f_X(x)\,dx + \int_{-\infty}^{\infty} cf_X(x)\,dx
\end{aligned}$$

by the additivity of the integral. Since a, b, and c are constants, they factor out of their respective integrals and we have

$$\begin{aligned}
&\mathrm{E}(ag_1(X)+bg_2(X)+c)\\
&\quad = a\int_{-\infty}^{\infty} g_1(x)f_X(x)\,dx + b\int_{-\infty}^{\infty} g_2(x)f_X(x)\,dx + c\int_{-\infty}^{\infty} f_X(x)\,dx\\
&\quad = a\mathrm{E}\,g_1(X) + b\mathrm{E}\,g_2(x) + c,
\end{aligned}$$

establishing (a). The other three properties are proved in a similar manner. □

Example 2.2.6 (Minimizing distance) The expected value of a random variable has another property, one that we can think of as relating to the interpretation of $\mathrm{E}\,X$ as a good guess at a value of X.

Suppose we measure the distance between a random variable X and a constant b by $(X-b)^2$. The closer b is to X, the smaller this quantity is. We can now determine the value of b that minimizes $\mathrm{E}(X-b)^2$ and, hence, will provide us with a good predictor of X. (Note that it does no good to look for a value of b that minimizes $(X-b)^2$, since the answer would depend on X, making it a useless predictor of X.)

We could proceed with the minimization of $\mathrm{E}(X-b)^2$ by using calculus, but there is a simpler method. (See Exercise 2.19 for a calculus-based proof.) Using the belief that there is something special about $\mathrm{E}\,X$, we write

$$\begin{aligned}
\mathrm{E}(X-b)^2 &= \mathrm{E}(X-\mathrm{E}\,X+\mathrm{E}\,X-b)^2 && \left(\begin{array}{c}\text{add } \pm\mathrm{E}\,X\text{, which}\\ \text{changes nothing}\end{array}\right)\\
&= \mathrm{E}\left((X-\mathrm{E}\,X)+(\mathrm{E}\,X-b)\right)^2 && \text{(group terms)}\\
&= \mathrm{E}(X-\mathrm{E}\,X)^2+(\mathrm{E}\,X-b)^2+2\mathrm{E}\left((X-\mathrm{E}\,X)(\mathrm{E}\,X-b)\right),
\end{aligned}$$

where we have expanded the square. Now, note that

$$\mathrm{E}\left((X-\mathrm{E}\,X)(\mathrm{E}\,X-b)\right)=(\mathrm{E}\,X-b)\mathrm{E}(X-\mathrm{E}\,X)=0,$$

since $(\mathrm{E}\,X-b)$ is constant and comes out of the expectation, and $\mathrm{E}(X-\mathrm{E}\,X)=\mathrm{E}\,X-\mathrm{E}\,X=0$. This means that

$$\mathrm{E}(X-b)^2=\mathrm{E}(X-\mathrm{E}\,X)^2+(\mathrm{E}\,X-b)^2. \tag{2.2.2}$$

We have no control over the first term on the right-hand side of (2.2.2), and the second term, which is always greater than or equal to 0, can be made equal to 0 by choosing $b=\mathrm{E}\,X$. Hence,

$$\min_b \mathrm{E}(X-b)^2=\mathrm{E}(X-\mathrm{E}\,X)^2. \tag{2.2.3}$$

See Exercise 2.18 for a similar result about the median. ||

When evaluating expectations of nonlinear functions of X, we can proceed in one of two ways. From the definition of $\mathrm{E}\,g(X)$, we could directly calculate

$$\mathrm{E}\,g(X)=\int_{-\infty}^{\infty} g(x)f_X(x)\,dx. \tag{2.2.4}$$

But we could also find the pdf $f_Y(y)$ of $Y=g(X)$ and we would have

$$\mathrm{E}\,g(X)=\mathrm{E}\,Y=\int_{-\infty}^{\infty} y f_Y(y)\,dy. \tag{2.2.5}$$

Example 2.2.7 (Uniform-exponential relationship–II) Let X have a uniform$(0,1)$ distribution, that is, the pdf of X is given by

$$f_X(x)=\begin{cases}1 & \text{if } 0\le x\le 1\\ 0 & \text{otherwise,}\end{cases}$$

and define a new random variable $g(X) = -\log X$. Then

$$\mathrm{E}\, g(X) = \mathrm{E}(-\log X) = \int_0^1 -\log x\, dx = x - x\log x\Big|_0^1 = 1.$$

But we also saw in Example 2.1.4 that $Y = -\log X$ has cdf $1 - e^{-y}$ and, hence, pdf $f_Y(y) = \frac{d}{dy}(1 - e^{-y}) = e^{-y}$, $0 < y < \infty$, which is a special case of the exponential pdf with $\lambda = 1$. Thus, by Example 2.2.2, $\mathrm{E}\, Y = 1$. ‖

2.3 Moments and Moment Generating Functions

The various moments of a distribution are an important class of expectations.

Definition 2.3.1 For each integer n, the nth *moment of* X (or $F_X(x)$), μ_n', is

$$\mu_n' = \mathrm{E}\, X^n.$$

The nth *central moment of* X, μ_n, is

$$\mu_n = \mathrm{E}(X - \mu)^n,$$

where $\mu = \mu_1' = \mathrm{E}\, X$.

Aside from the mean, $\mathrm{E}\, X$, of a random variable, perhaps the most important moment is the second central moment, more commonly known as the variance.

Definition 2.3.2 The *variance* of a random variable X is its second central moment, $\mathrm{Var}\, X = \mathrm{E}(X - \mathrm{E}\, X)^2$. The positive square root of $\mathrm{Var}\, X$ is the *standard deviation* of X.

The variance gives a measure of the degree of spread of a distribution around its mean. We saw earlier in Example 2.2.6 that the quantity $\mathrm{E}(X - b)^2$ is minimized by choosing $b = \mathrm{E}\, X$. Now we consider the absolute size of this minimum. The interpretation attached to the variance is that larger values mean X is more variable. At the extreme, if $\mathrm{Var}\, X = \mathrm{E}(X - \mathrm{E}\, X)^2 = 0$, then X is equal to $\mathrm{E}\, X$ with probability 1, and there is no variation in X. The standard deviation has the same qualitative interpretation: Small values mean X is very likely to be close to $\mathrm{E}\, X$, and large values mean X is very variable. The standard deviation is easier to interpret in that the measurement unit on the standard deviation is the same as that for the original variable X. The measurement unit on the variance is the square of the original unit.

Example 2.3.3 (Exponential variance) Let X have the exponential(λ) distribution, defined in Example 2.2.2. There we calculated $\mathrm{E}X = \lambda$, and we can now calculate the variance by

$$\begin{aligned}\mathrm{Var}\, X = \mathrm{E}(X - \lambda)^2 &= \int_0^\infty (x - \lambda)^2 \frac{1}{\lambda} e^{-x/\lambda}\, dx \\ &= \int_0^\infty (x^2 - 2x\lambda + \lambda^2) \frac{1}{\lambda} e^{-x/\lambda}\, dx.\end{aligned}$$

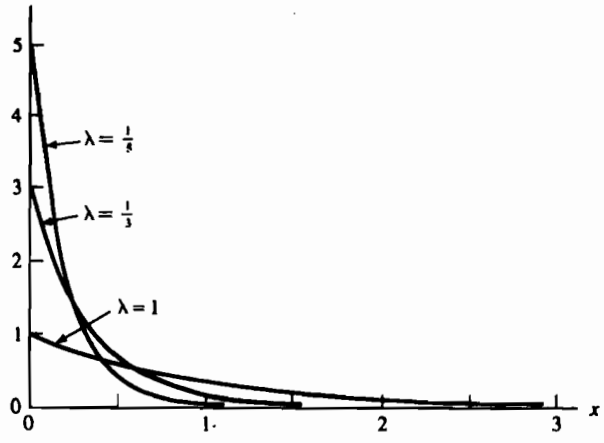


Figure 2.3.1. *Exponential densities for* $\lambda = 1, \frac{1}{3}, \frac{1}{5}$

To complete the integration, we can integrate each of the terms separately, using integration by parts on the terms involving x and x^2. Upon doing this, we find that $\text{Var } X = \lambda^2$. ‖

We see that the variance of an exponential distribution is directly related to the parameter λ. Figure 2.3.1 shows several exponential distributions corresponding to different values of λ. Notice how the distribution is more concentrated about its mean for smaller values of λ. The behavior of the variance of an exponential, as a function of λ, is a special case of the variance behavior summarized in the following theorem.

Theorem 2.3.4 *If X is a random variable with finite variance, then for any constants a and b,*

$$\text{Var}(aX + b) = a^2 \text{ Var } X.$$

Proof: From the definition, we have

$$\begin{aligned} \text{Var}(aX + b) &= \text{E}\left((aX + b) - \text{E}(aX + b)\right)^2 \\ &= \text{E}(aX - a\text{E}\,X)^2 \qquad (\text{E}(aX + b) = a\text{E}\,X + b) \\ &= a^2\text{E}(X - \text{E}\,X)^2 \\ &= a^2 \text{ Var } X. \end{aligned}$$

□

It is sometimes easier to use an alternative formula for the variance, given by

$$\text{Var } X = \text{E}\,X^2 - (\text{E}\,X)^2, \tag{2.3.1}$$

which is easily established by noting

$$\begin{aligned}\operatorname{Var} X = \mathrm{E}(X - \mathrm{E}X)^2 &= \mathrm{E}[X^2 - 2X\mathrm{E}X + (\mathrm{E}X)^2] \\ &= \mathrm{E}X^2 - 2(\mathrm{E}X)^2 + (\mathrm{E}X)^2 \\ &= \mathrm{E}X^2 - (\mathrm{E}X)^2,\end{aligned}$$

where we use the fact that $\mathrm{E}(X\mathrm{E}X) = (\mathrm{E}X)(\mathrm{E}X) = (\mathrm{E}X)^2$, since $\mathrm{E}X$ is a constant. We now illustrate some moment calculations with a discrete distribution.

Example 2.3.5 (Binomial variance) Let $X \sim \text{binomial}(n, p)$, that is,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

We have previously seen that $\mathrm{E}X = np$. To calculate $\operatorname{Var} X$ we first calculate $\mathrm{E}X^2$. We have

$$\mathrm{E}X^2 = \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x}. \tag{2.3.2}$$

In order to sum this series, we must first manipulate the binomial coefficient in a manner similar to that used for $\mathrm{E}X$ (Example 2.2.3). We write

$$x^2 \binom{n}{x} = x \frac{n!}{(x-1)!(n-x)!} = xn \binom{n-1}{x-1}. \tag{2.3.3}$$

The summand in (2.3.2) corresponding to $x = 0$ is 0, and using (2.3.3), we have

$$\begin{aligned}\mathrm{E}X^2 &= n \sum_{x=1}^{n} x \binom{n-1}{x-1} p^x (1-p)^{n-x} \\ &= n \sum_{y=0}^{n-1} (y+1) \binom{n-1}{y} p^{y+1} (1-p)^{n-1-y} \qquad \text{(setting } y = x-1\text{)} \\ &= np \sum_{y=0}^{n-1} y \binom{n-1}{y} p^y (1-p)^{n-1-y} + np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-1-y}.\end{aligned}$$

Now it is easy to see that the first sum is equal to $(n-1)p$ (since it is the mean of a binomial$(n-1, p)$), while the second sum is equal to 1. Hence,

$$\mathrm{E}X^2 = n(n-1)p^2 + np. \tag{2.3.4}$$

Using (2.3.1), we have

$$\operatorname{Var} X = n(n-1)p^2 + np - (np)^2 = -np^2 + np = np(1-p). \qquad \|$$

Calculation of higher moments proceeds in an analogous manner, but usually the mathematical manipulations become quite involved. In applications, moments of order 3 or 4 are sometimes of interest, but there is usually little statistical reason for examining higher moments than these.

We now introduce a new function that is associated with a probability distribution, the *moment generating function* (mgf). As its name suggests, the mgf can be used to generate moments. In practice, it is easier in many cases to calculate moments directly than to use the mgf. However, the main use of the mgf is not to generate moments, but to help in characterizing a distribution. This property can lead to some extremely powerful results when used properly.

Definition 2.3.6 Let X be a random variable with cdf F_X. The *moment generating function (mgf) of* X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = \mathrm{E}\, e^{tX},$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, $\mathrm{E}e^{tX}$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \qquad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x e^{tx} P(X = x) \qquad \text{if } X \text{ is discrete.}$$

It is very easy to see how the mgf generates moments. We summarize the result in the following theorem.

Theorem 2.3.7 *If X has mgf $M_X(t)$, then*

$$\mathrm{E}\, X^n = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\bigg|_{t=0}.$$

That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$.

Proof: Assuming that we can differentiate under the integral sign (see the next section), we have

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \left(\frac{d}{dt} e^{tx} \right) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} (x e^{tx}) f_X(x)\, dx \\
&= \mathrm{E}\, X e^{tX}.
\end{aligned}$$

Thus,

$$\frac{d}{dt}M_X(t)\bigg|_{t=0} = \mathrm{E}\,Xe^{tX}\big|_{t=0} = \mathrm{E}\,X.$$

Proceeding in an analogous manner, we can establish that

$$\frac{d^n}{dt^n}M_X(t)\bigg|_{t=0} = \mathrm{E}\,X^n e^{tX}\big|_{t=0} = \mathrm{E}\,X^n. \qquad \square$$

Example 2.3.8 (Gamma mgf) In Example 2.1.6 we encountered a special case of the gamma pdf,

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

where $\Gamma(\alpha)$ denotes the gamma function, some of whose properties are given in Section 3.3. The mgf is given by

$$\begin{aligned} M_X(t) &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty e^{tx}x^{\alpha-1}e^{-x/\beta}\,dx \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-(\frac{1}{\beta}-t)x}\,dx \qquad (2.3.5) \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha}\int_0^\infty x^{\alpha-1}e^{-x/(\frac{\beta}{1-\beta t})}\,dx. \end{aligned}$$

We now recognize the integrand in (2.3.5) as the *kernel* of another gamma pdf. (The *kernel* of a function is the main part of the function, the part that remains when constants are disregarded.) Using the fact that, for any positive constants a and b,

$$f(x) = \frac{1}{\Gamma(a)b^a}x^{a-1}e^{-x/b}$$

is a pdf, we have that

$$\int_0^\infty \frac{1}{\Gamma(a)b^a}x^{a-1}e^{-x/b}\,dx = 1$$

and, hence,

$$\int_0^\infty x^{a-1}e^{-x/b}\,dx = \Gamma(a)b^a. \qquad (2.3.6)$$

Applying (2.3.6) to (2.3.5), we have

$$M_X(t) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\Gamma(\alpha)\left(\frac{\beta}{1-\beta t}\right)^\alpha = \left(\frac{1}{1-\beta t}\right)^\alpha \qquad \text{if } t < \frac{1}{\beta}.$$

If $t \geq 1/\beta$, then the quantity $(1/\beta) - t$, in the integrand of (2.3.5), is nonpositive and the integral in (2.3.6) is infinite. Thus, the mgf of the gamma distribution exists only if $t < 1/\beta$. (In Section 3.3 we will explore the gamma function in more detail.)

The mean of the gamma distribution is given by

$$\mathrm{E}\,X = \frac{d}{dt} M_X(t)\Big|_{t=0} = \frac{\alpha\beta}{(1-\beta t)^{\alpha+1}}\Big|_{t=0} = \alpha\beta.$$

Other moments can be calculated in a similar manner. ||

Example 2.3.9 (Binomial mgf) For a second illustration of calculating a moment generating function, we consider a discrete distribution, the binomial distribution. The binomial(n, p) pmf is given in (2.1.3). So

$$M_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x}.$$

The binomial formula (see Theorem 3.2.2) gives

$$\sum_{x=0}^{n} \binom{n}{x} u^x v^{n-x} = (u+v)^n. \tag{2.3.7}$$

Hence, letting $u = pe^t$ and $v = 1 - p$, we have

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

As previously mentioned, the major usefulness of the moment generating function is not in its ability to generate moments. Rather, its usefulness stems from the fact that, in many cases, the moment generating function can characterize a distribution. There are, however, some technical difficulties associated with using moments to characterize a distribution, which we will now investigate.

If the mgf exists, it characterizes an infinite set of moments. The natural question is whether characterizing the infinite set of moments uniquely determines a distribution function. The answer to this question, unfortunately, is no. Characterizing the set of moments is not enough to determine a distribution uniquely because there may be two distinct random variables having the same moments.

Example 2.3.10 (Nonunique moments) Consider the two pdfs given by

$$f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}, \qquad 0 \leq x < \infty,$$

$$f_2(x) = f_1(x)[1 + \sin(2\pi \log x)], \qquad 0 \leq x < \infty.$$

(The pdf f_1 is a special case of a *lognormal pdf.*)

It can be shown that if $X_1 \sim f_1(x)$, then

$$\mathrm{E}\,X_1^r = e^{r^2/2}, \qquad r = 0, 1, \ldots,$$

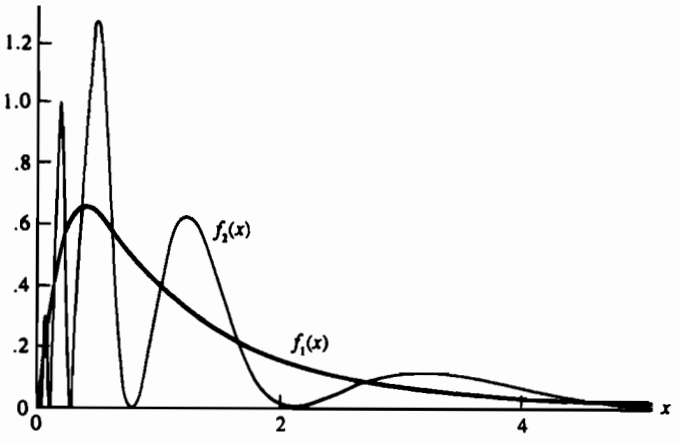


Figure 2.3.2. *Two pdfs with the same moments:* $f_1(x) = \frac{1}{\sqrt{2\pi}x} e^{-(\log x)^2/2}$ *and* $f_2(x) = f_1(x)[1 + \sin(2\pi \log x)]$

so X_1 has all of its moments. Now suppose that $X_2 \sim f_2(x)$. We have

$$\begin{aligned} \mathrm{E}\, X_2^r &= \int_0^\infty x^r f_1(x)[1 + \sin(2\pi \log x)]\, dx \\ &= \mathrm{E}\, X_1^r + \int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx. \end{aligned}$$

However, the transformation $y = \log x - r$ shows that this last integral is that of an odd function over $(-\infty, \infty)$ and hence is equal to 0 for $r = 0, 1, \ldots$. Thus, even though X_1 and X_2 have distinct pdfs, they have the same moments for all r. The two pdfs are pictured in Figure 2.3.2.

See Exercise 2.35 for details and also Exercises 2.34, 2.36, and 2.37 for more about mgfs and distributions. ||

The problem of uniqueness of moments does not occur if the random variables have bounded support. If that is the case, then the infinite sequence of moments does uniquely determine the distribution (see, for example, Billingsley 1995, Section 30). Furthermore, if the mgf exists in a neighborhood of 0, then the distribution *is* uniquely determined, no matter what its support. Thus, existence of all moments is not equivalent to existence of the moment generating function. The following theorem shows how a distribution can be characterized.

Theorem 2.3.11 *Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.*

a. *If X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if $\mathrm{E}\, X^r = \mathrm{E}\, Y^r$ for all integers $r = 0, 1, 2, \ldots$.*

b. *If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.*

In the next theorem, which deals with a sequence of mgfs that converges, we do not treat the bounded support case separately. Note that the uniqueness assumption is automatically satisfied if the limiting mgf exists in a neighborhood of 0 (see Miscellanea 2.6.1).

Theorem 2.3.12 (Convergence of mgfs) *Suppose $\{X_i, i = 1, 2, \ldots\}$ is a sequence of random variables, each with mgf $M_{X_i}(t)$. Furthermore, suppose that*

$$\lim_{i \to \infty} M_{X_i}(t) = M_X(t), \qquad \text{for all } t \text{ in a neighborhood of } 0,$$

and $M_X(t)$ is an mgf. Then there is a unique cdf F_X whose moments are determined by $M_X(t)$ and, for all x where $F_X(x)$ is continuous, we have

$$\lim_{i \to \infty} F_{X_i}(x) = F_X(x).$$

That is, convergence, for $|t| < h$, of mgfs to an mgf implies convergence of cdfs.

The proofs of Theorems 2.3.11 and 2.3.12 rely on the theory of *Laplace transforms*. (The classic reference is Widder 1946, but Laplace transforms also get a comprehensive treatment by Feller 1971.) The defining equation for $M_X(t)$, that is,

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx, \tag{2.3.8}$$

defines a Laplace transform ($M_X(t)$ is the Laplace transform of $f_X(x)$). A key fact about Laplace transforms is their uniqueness. If (2.3.8) is valid for all t such that $|t| < h$, where h is some positive number, then given $M_X(t)$ there is only one function $f_X(x)$ that satisfies (2.3.8). Given this fact, the two previous theorems are quite reasonable. While rigorous proofs of these theorems are not beyond the scope of this book, the proofs are technical in nature and provide no real understanding. We omit them.

The possible nonuniqueness of the moment sequence is an annoyance. If we show that a sequence of moments converges, we will not be able to conclude formally that the random variables converge. To do so, we would have to verify the uniqueness of the moment sequence, a generally horrible job (see Miscellanea 2.6.1). However, if the sequence of mgfs converges in a neighborhood of 0, then the random variables converge. Thus, we can consider the convergence of mgfs as a sufficient, but not necessary, condition for the sequence of random variables to converge.

Example 2.3.13 (Poisson approximation) One approximation that is usually taught in elementary statistics courses is that binomial probabilities (see Example 2.3.5) can be approximated by *Poisson* probabilities, which are generally easier to calculate. The binomial distribution is characterized by two quantities, denoted by n and p. It is taught that the Poisson approximation is valid "when n is large and np is small," and rules of thumb are sometimes given.

The Poisson(λ) pmf is given by

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

where λ is a positive constant. The approximation states that if $X \sim \text{binomial}(n, p)$ and $Y \sim \text{Poisson}(\lambda)$, with $\lambda = np$, then

$$P(X = x) \approx P(Y = x) \tag{2.3.9}$$

for large n and small np. We now show that the mgfs converge, lending credence to this approximation. Recall that

$$M_X(t) = [pe^t + (1-p)]^n. \tag{2.3.10}$$

For the Poisson(λ) distribution, we can calculate (see Exercise 2.33)

$$M_Y(t) = e^{\lambda(e^t - 1)},$$

and if we define $p = \lambda/n$, then $M_X(t) \to M_Y(t)$ as $n \to \infty$. The validity of the approximation in (2.3.9) will then follow from Theorem 2.3.12.

We first must digress a bit and mention an important limit result, one that has wide applicability in statistics. The proof of this lemma may be found in many standard calculus texts.

Lemma 2.3.14 *Let $a_1, a_2, \ldots$ be a sequence of numbers converging to a, that is, $\lim_{n\to\infty} a_n = a$. Then*

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Returning to the example, we have

$$M_X(t) = [pe^t + (1-p)]^n = \left[1 + \frac{1}{n}(e^t - 1)(np)\right]^n = \left[1 + \frac{1}{n}(e^t - 1)\lambda\right]^n,$$

because $\lambda = np$. Now set $a_n = a = (e^t - 1)\lambda$, and apply Lemma 2.3.14 to get

$$\lim_{n\to\infty} M_X(t) = e^{\lambda(e^t - 1)} = M_Y(t),$$

the moment generating function of the Poisson.

The Poisson approximation can be quite good even for moderate p and n. In Figure 2.3.3 we show a binomial mass function along with its Poisson approximation, with $\lambda = np$. The approximation appears to be satisfactory. ||

We close this section with a useful result concerning mgfs.

Theorem 2.3.15 *For any constants a and b, the mgf of the random variable $aX + b$ is given by*

$$M_{aX+b}(t) = e^{bt} M_X(at).$$

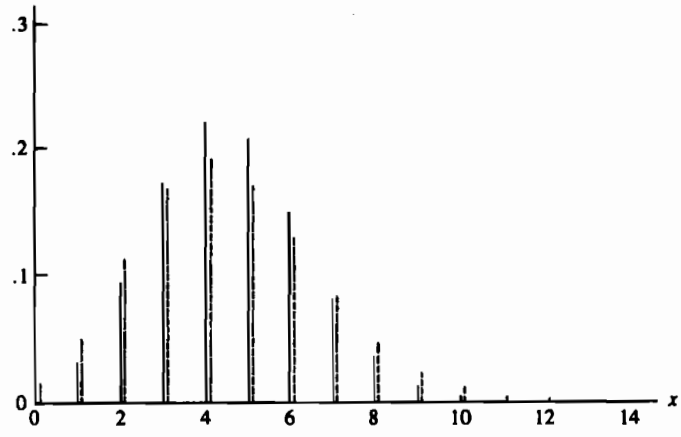


Figure 2.3.3. *Poisson (dotted line) approximation to the binomial (solid line), $n = 15$, $p = .3$*

Proof: By definition,

$$\begin{aligned} M_{aX+b}(t) &= \mathrm{E}\left(e^{(aX+b)t}\right) \\ &= \mathrm{E}\left(e^{(aX)t}e^{bt}\right) && \text{(properties of exponentials)} \\ &= e^{bt}\mathrm{E}\left(e^{(at)X}\right) && (e^{bt} \text{ is constant}) \\ &= e^{bt}M_X(at), && \text{(definition of mgf)} \end{aligned}$$

proving the theorem. □

2.4 Differentiating Under an Integral Sign

In the previous section we encountered an instance in which we desired to interchange the order of integration and differentiation. This situation is encountered frequently in theoretical statistics. The purpose of this section is to characterize conditions under which this operation is legitimate. We will also discuss interchanging the order of differentiation and summation.

Many of these conditions can be established using standard theorems from calculus and detailed proofs can be found in most calculus textbooks. Thus, detailed proofs will not be presented here.

We first want to establish the method of calculating

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,dx, \tag{2.4.1}$$

where $-\infty < a(\theta), b(\theta) < \infty$ for all θ. The rule for differentiating (2.4.1) is called Leibnitz's Rule and is an application of the Fundamental Theorem of Calculus and the chain rule.

Theorem 2.4.1 (Leibnitz's Rule) *If $f(x,\theta)$, $a(\theta)$, and $b(\theta)$ are differentiable with respect to θ, then*

$$\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)\,dx = f(b(\theta),\theta)\frac{d}{d\theta}b(\theta) - f(a(\theta),\theta)\frac{d}{d\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f(x,\theta)\,dx.$$

Notice that if $a(\theta)$ and $b(\theta)$ are constant, we have a special case of Leibnitz's Rule:

$$\frac{d}{d\theta}\int_a^b f(x,\theta)\,dx = \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)\,dx.$$

Thus, in general, if we have the integral of a differentiable function over a finite range, differentiation of the integral poses no problem. If the range of integration is infinite, however, problems can arise.

Note that the interchange of derivative and integral in the above equation equates a partial derivative with an ordinary derivative. Formally, this must be the case since the left-hand side is a function of only θ, while the integrand on the right-hand side is a function of both θ and x.

The question of whether interchanging the order of differentiation and integration is justified is really a question of whether limits and integration can be interchanged, since a derivative is a special kind of limit. Recall that if $f(x,\theta)$ is differentiable, then

$$\frac{\partial}{\partial\theta} f(x,\theta) = \lim_{\delta\to 0}\frac{f(x,\theta+\delta) - f(x,\theta)}{\delta},$$

so we have

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta} f(x,\theta)\,dx = \int_{-\infty}^{\infty}\lim_{\delta\to 0}\left[\frac{f(x,\theta+\delta) - f(x,\theta)}{\delta}\right]dx,$$

while

$$\frac{d}{d\theta}\int_{-\infty}^{\infty} f(x,\theta)\,dx = \lim_{\delta\to 0}\int_{-\infty}^{\infty}\left[\frac{f(x,\theta+\delta) - f(x,\theta)}{\delta}\right]dx.$$

Therefore, if we can justify the interchanging of the order of limits and integration, differentiation under the integral sign will be justified. Treatment of this problem in full generality will, unfortunately, necessitate the use of measure theory, a topic that will not be covered in this book. However, the statements and conclusions of some important results can be given. The following theorems are all corollaries of Lebesgue's Dominated Convergence Theorem (see, for example, Rudin 1976).

Theorem 2.4.2 *Suppose the function $h(x,y)$ is continuous at y_0 for each x, and there exists a function $g(x)$ satisfying*

i. $|h(x,y)| \le g(x)$ *for all x and y,*

ii. $\int_{-\infty}^{\infty} g(x)\,dx < \infty$.

Then

$$\lim_{y\to y_0}\int_{-\infty}^{\infty} h(x,y)\,dx = \int_{-\infty}^{\infty}\lim_{y\to y_0} h(x,y)\,dx.$$

The key condition in this theorem is the existence of a dominating function $g(x)$, with a finite integral, which ensures that the integrals cannot be too badly behaved. We can now apply this theorem to the case we are considering by identifying $h(x, y)$ with the difference $(f(x, \theta + \delta) - f(x, \theta))/\delta$.

Theorem 2.4.3 *Suppose $f(x, \theta)$ is differentiable at $\theta = \theta_0$, that is,*

$$\lim_{\delta \to 0} \frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} = \frac{\partial}{\partial \theta} f(x, \theta)\bigg|_{\theta = \theta_0}$$

exists for every x, and there exists a function $g(x, \theta_0)$ and a constant $\delta_0 > 0$ such that

i. $\left| \dfrac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} \right| \leq g(x, \theta_0)$, *for all x and $|\delta| \leq \delta_0$,*

ii. $\int_{-\infty}^{\infty} g(x, \theta_0)\, dx < \infty$.

Then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)\, dx\bigg|_{\theta = \theta_0} = \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial \theta} f(x, \theta)\bigg|_{\theta = \theta_0} \right] dx. \tag{2.4.2}$$

Condition (i) is similar to what is known as a *Lipschitz condition*, a condition that imposes smoothness on a function. Here, condition (i) is effectively bounding the variability in the first derivative; other smoothness constraints might bound this variability by a constant (instead of a function g), or place a bound on the variability of the second derivative of f.

The conclusion of Theorem 2.4.3 is a little cumbersome, but it is important to realize that although we seem to be treating θ as a variable, the statement of the theorem is for one value of θ. That is, for each value θ_0 for which $f(x, \theta)$ is differentiable at θ_0 and satisfies conditions (i) and (ii), the order of integration and differentiation can be interchanged. Often the distinction between θ and θ_0 is not stressed and (2.4.2) is written

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)\, dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta)\, dx. \tag{2.4.3}$$

Typically, $f(x, \theta)$ is differentiable at all θ, not at just one value θ_0. In this case, condition (i) of Theorem 2.4.3 can be replaced by another condition that often proves easier to verify. By an application of the mean value theorem, it follows that, for fixed x and θ_0, and $|\delta| \leq \delta_0$,

$$\frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} = \frac{\partial}{\partial \theta} f(x, \theta)\bigg|_{\theta = \theta_0 + \delta^*(x)}$$

for some number $\delta^*(x)$, where $|\delta^*(x)| \leq \delta_0$. Therefore, condition (i) will be satisfied if we find a $g(x, \theta)$ that satisfies condition (ii) and

$$\left| \frac{\partial}{\partial \theta} f(x, \theta)\bigg|_{\theta = \theta'} \right| \leq g(x, \theta) \qquad \text{for all } \theta' \text{ such that } |\theta' - \theta| \leq \delta_0. \tag{2.4.4}$$

Note that in (2.4.4) δ_0 is implicitly a function of θ, as is the case in Theorem 2.4.3. This is permitted since the theorem is applied to each value of θ individually. From (2.4.4) we get the following corollary.

Corollary 2.4.4 *Suppose $f(x,\theta)$ is differentiable in θ and there exists a function $g(x,\theta)$ such that (2.4.4) is satisfied and $\int_{-\infty}^{\infty} g(x,\theta)\,dx < \infty$. Then (2.4.3) holds.*

Notice that both condition (i) of Theorem 2.4.3 and (2.4.4) impose a uniformity requirement on the functions to be bounded; some type of uniformity is generally needed before derivatives and integrals can be interchanged.

Example 2.4.5 (Interchanging integration and differentiation–I) Let X have the exponential(λ) pdf given by $f(x) = (1/\lambda)e^{-x/\lambda}$, $0 < x < \infty$, and suppose we want to calculate

$$\frac{d}{d\lambda}\,\mathrm{E}\,X^n = \frac{d}{d\lambda}\int_0^\infty x^n \left(\frac{1}{\lambda}\right) e^{-x/\lambda}\,dx \tag{2.4.5}$$

for integer $n > 0$. If we could move the differentiation inside the integral, we would have

$$\begin{aligned}\frac{d}{d\lambda}\,\mathrm{E}\,X^n &= \int_0^\infty \frac{\partial}{\partial\lambda} x^n \left(\frac{1}{\lambda}\right) e^{-x/\lambda}\,dx \\ &= \int_0^\infty \frac{x^n}{\lambda^2}\left(\frac{x}{\lambda} - 1\right) e^{-x/\lambda}\,dx \qquad (2.4.6)\\ &= \frac{1}{\lambda^2}\,\mathrm{E}\,X^{n+1} - \frac{1}{\lambda}\,\mathrm{E}\,X^n.\end{aligned}$$

To justify the interchange of integration and differentiation, we bound the derivative of $x^n(1/\lambda)e^{-x/\lambda}$. Now

$$\left|\frac{\partial}{\partial\lambda}\left(\frac{x^n e^{-x\lambda}}{\lambda}\right)\right| = \frac{x^n e^{-x/\lambda}}{\lambda^2}\left|\frac{x}{\lambda} - 1\right| \le \frac{x^n e^{-x/\lambda}}{\lambda^2}\left(\frac{x}{\lambda} + 1\right). \qquad (\text{since } \tfrac{x}{\lambda} > 0)$$

For some constant δ_0 satisfying $0 < \delta_0 < \lambda$, take

$$g(x,\lambda) = \frac{x^n e^{-x/(\lambda+\delta_0)}}{(\lambda-\delta_0)^2}\left(\frac{x}{\lambda-\delta_0} + 1\right).$$

We then have

$$\left|\frac{\partial}{\partial\lambda}\left(\frac{x^n e^{-x/\lambda}}{\lambda}\right)\Bigg|_{\lambda=\lambda'}\right| \le g(x,\lambda) \qquad \text{for all } \lambda' \text{ such that } |\lambda' - \lambda| \le \delta_0.$$

Since the exponential distribution has all of its moments, $\int_{-\infty}^{\infty} g(x,\lambda)\,dx < \infty$ as long as $\lambda - \delta_0 > 0$, so the interchange of integration and differentiation is justified. ‖

The property illustrated for the exponential distribution holds for a large class of densities, which will be dealt with in Section 3.4.

Notice that (2.4.6) gives us a recursion relation for the moments of the exponential distribution,

$$\mathrm{E}\,X^{n+1} = \lambda \mathrm{E}\,X^n + \lambda^2 \frac{d}{d\lambda} \mathrm{E}\,X^n, \tag{2.4.7}$$

making the calculation of the $(n+1)$st moment relatively easy. This type of relationship exists for other distributions. In particular, if X has a normal distribution with mean μ and variance 1, so it has pdf $f(x) = (1/\sqrt{2\pi})e^{-(x-\mu)^2/2}$, then

$$\mathrm{E}\,X^{n+1} = \mu \mathrm{E}\,X^n - \frac{d}{d\mu} \mathrm{E}\,X^n.$$

We illustrate one more interchange of differentiation and integration, one involving the moment generating function.

Example 2.4.6 (Interchanging integration and differentiation–II) Again let X have a normal distribution with mean μ and variance 1, and consider the mgf of X,

$$M_X(t) = \mathrm{E}\,e^{tX} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2}\,dx.$$

In Section 2.3 it was stated that we can calculate moments by differentiation of $M_X(t)$ and differentiation under the integral sign was justified:

$$\frac{d}{dt} M_X(t) = \frac{d}{dt} \mathrm{E}\,e^{tX} = \mathrm{E} \frac{\partial}{\partial t} e^{tX} = \mathrm{E}(Xe^{tX}). \tag{2.4.8}$$

We can apply the results of this section to justify the operations in (2.4.8). Notice that when applying either Theorem 2.4.3 or Corollary 2.4.4 here, we identify t with the variable θ in Theorem 2.4.3. The parameter μ is treated as a constant.

From Corollary 2.4.4, we must find a function $g(x,t)$, with finite integral, that satisfies

$$\left.\frac{\partial}{\partial t} e^{tx} e^{-(x-\mu)^2/2}\right|_{t=t'} \le g(x,t) \quad \text{for all } t' \text{ such that } |t'-t| \le \delta_0. \tag{2.4.9}$$

Doing the obvious, we have

$$\left|\frac{\partial}{\partial t} e^{tx} e^{-(x-\mu)^2/2}\right| = \left|xe^{tx} e^{-(x-\mu)^2/2}\right| \le |x| e^{tx} e^{-(x-\mu)^2/2}.$$

It is easiest to define our function $g(x,t)$ separately for $x \ge 0$ and $x < 0$. We take

$$g(x,t) = \begin{cases} |x| e^{(t-\delta_0)x} e^{-(x-\mu)^2/2} & \text{if } x < 0 \\ |x| e^{(t+\delta_0)x} e^{-(x-\mu)^2/2} & \text{if } x \ge 0. \end{cases}$$

It is clear that this function satisfies (2.4.9); it remains to check that its integral is finite.

For $x \geq 0$ we have

$$g(x,t) = xe^{-(x^2-2x(\mu+t+\delta_0)+\mu^2)/2}.$$

We now complete the square in the exponent; that is, we write

$$\begin{aligned} x^2 - 2x(\mu+t+\delta_0) + \mu^2 \\ &= x^2 - 2x(\mu+t+\delta_0) + (\mu+t+\delta_0)^2 - (\mu+t+\delta_0)^2 + \mu^2 \\ &= (x-(\mu+t+\delta_0))^2 + \mu^2 - (\mu+t+\delta_0)^2, \end{aligned}$$

and so, for $x \geq 0$,

$$g(x,t) = xe^{-[x-(\mu+t+\delta_0)]^2/2}e^{-[\mu^2-(\mu+t+\delta_0)^2]/2}.$$

Since the last exponential factor in this expression does not depend on x, $\int_0^\infty g(x,t)\,dx$ is essentially calculating the mean of a normal distribution with mean $\mu+t+\delta_0$, except that the integration is only over $[0,\infty)$. However, it follows that the integral is finite because the normal distribution has a finite mean (to be shown in Chapter 3). A similar development for $x < 0$ shows that

$$g(x,t) = |x|e^{-[x-(\mu+t-\delta_0)]^2/2}e^{-[\mu^2-(\mu+t-\delta_0)^2]/2}$$

and so $\int_{-\infty}^0 g(x,t)\,dx < \infty$. Therefore, we have found an integrable function satisfying (2.4.9) and the operation in (2.4.8) is justified. ||

We now turn to the question of when it is possible to interchange differentiation and summation, an operation that plays an important role in discrete distributions. Of course, we are concerned only with infinite sums, since a derivative can always be taken inside a finite sum.

Example 2.4.7 (Interchanging summation and differentiation) Let X be a discrete random variable with the *geometric distribution*

$$P(X=x) = \theta(1-\theta)^x, \quad x = 0,1,\ldots, \quad 0 < \theta < 1.$$

We have that $\sum_{x=0}^\infty \theta(1-\theta)^x = 1$ and, provided that the operations are justified,

$$\begin{aligned} \frac{d}{d\theta}\sum_{x=0}^\infty \theta(1-\theta)^x &= \sum_{x=0}^\infty \frac{d}{d\theta}\theta(1-\theta)^x \\ &= \sum_{x=0}^\infty \left[(1-\theta)^x - \theta x(1-\theta)^{x-1}\right] \\ &= \frac{1}{\theta}\sum_{x=0}^\infty \theta(1-\theta)^x - \frac{1}{1-\theta}\sum_{x=0}^\infty x\theta(1-\theta)^x. \end{aligned}$$

Since $\sum_{x=0}^\infty \theta(1-\theta)^x = 1$ for all $0 < \theta < 1$, its derivative is 0. So we have

$$\frac{1}{\theta}\sum_{x=0}^\infty \theta(1-\theta)^x - \frac{1}{1-\theta}\sum_{x=0}^\infty x\theta(1-\theta)^x = 0. \tag{2.4.10}$$

Now the first sum in (2.4.10) is equal to 1 and the second sum is $\mathrm{E}\,X$; hence (2.4.10) becomes

$$\frac{1}{\theta} - \frac{1}{1-\theta}\,\mathrm{E}\,X = 0,$$

or

$$\mathrm{E}\,X = \frac{1-\theta}{\theta}.$$

We have, in essence, summed the series $\sum_{x=0}^{\infty} x\theta(1-\theta)^x$ by differentiating. ||

Justification for taking the derivative inside the summation is more straightforward than the integration case. The following theorem provides the details.

Theorem 2.4.8 *Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges for all θ in an interval (a, b) of real numbers and*

i. $\frac{\partial}{\partial\theta} h(\theta, x)$ *is continuous in θ for each x,*

ii. $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x)$ *converges uniformly on every closed bounded subinterval of (a, b).*

Then

$$\frac{d}{d\theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x). \tag{2.4.11}$$

The condition of uniform convergence is the key one to verify in order to establish that the differentiation can be taken inside the summation. Recall that a series converges uniformly if its sequence of partial sums converges uniformly, a fact that we use in the following example.

Example 2.4.9 (Continuation of Example 2.4.7) To apply Theorem 2.4.8 we identify

$$h(\theta, x) = \theta(1-\theta)^x$$

and

$$\frac{\partial}{\partial\theta} h(\theta, x) = (1-\theta)^x - \theta x(1-\theta)^{x-1},$$

and verify that $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(\theta, x)$ converges uniformly. Define $S_n(\theta)$ by

$$S_n(\theta) = \sum_{x=0}^{n} \left[(1-\theta)^x - \theta x(1-\theta)^{x-1}\right].$$

The convergence will be uniform on $[c, d] \subset (0, 1)$ if, given $\epsilon > 0$, we can find an N such that

$$n > N \Rightarrow |S_n(\theta) - S_\infty(\theta)| < \epsilon \quad \text{for all } \theta \in [c, d].$$

Recall the partial sum of the geometric series (1.5.3). If $y \neq 1$, then we can write

$$\sum_{k=0}^{n} y^k = \frac{1-y^{n+1}}{1-y}.$$

Applying this, we have

$$\begin{aligned}
\sum_{x=0}^{n}(1-\theta)^x &= \frac{1-(1-\theta)^{n+1}}{\theta} \\
\sum_{x=0}^{n}\theta x(1-\theta)^{x-1} &= \theta\sum_{x=0}^{n} -\frac{\partial}{\partial\theta}(1-\theta)^x \\
&= -\theta\frac{d}{d\theta}\sum_{x=0}^{n}(1-\theta)^x \\
&= -\theta\frac{d}{d\theta}\left[\frac{1-(1-\theta)^{n+1}}{\theta}\right].
\end{aligned}$$

Here we (justifiably) pull the derivative through the finite sum. Calculating this derivative gives

$$\sum_{x=0}^{n}\theta x(1-\theta)^{x-1} = \frac{(1-(1-\theta)^{n+1}) - (n+1)\theta(1-\theta)^n}{\theta},$$

and, hence,

$$\begin{aligned}
S_n(\theta) &= \frac{1-(1-\theta)^{n+1}}{\theta} - \frac{(1-(1-\theta)^{n+1}) - (n+1)\theta(1-\theta)^n}{\theta} \\
&= (n+1)(1-\theta)^n.
\end{aligned}$$

It is clear that, for $0 < \theta < 1$, $S_\infty = \lim_{n\to\infty} S_n(\theta) = 0$. Since $S_n(\theta)$ is continuous, the convergence is uniform on any closed bounded interval. Therefore, the series of derivatives converges uniformly and the interchange of differentiation and summation is justified. ||

We close this section with a theorem that is similar to Theorem 2.4.8, but treats the case of interchanging the order of summation and integration.

Theorem 2.4.10 *Suppose the series $\sum_{x=0}^{\infty} h(\theta, x)$ converges uniformly on $[a, b]$ and that, for each x, $h(\theta, x)$ is a continuous function of θ. Then*

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x)\, d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x)\, d\theta.$$

2.5 Exercises

2.1 In each of the following find the pdf of Y. Show that the pdf integrates to 1.

(a) $Y = X^3$ and $f_X(x) = 42x^5(1-x)$, $0 < x < 1$

(b) $Y = 4X + 3$ and $f_X(x) = 7e^{-7x}$, $0 < x < \infty$

(c) $Y = X^2$ and $f_X(x) = 30x^2(1-x)^2$, $0 < x < 1$

(See Example A.0.2 in Appendix A.)

2.2 In each of the following find the pdf of Y.

(a) $Y = X^2$ and $f_X(x) = 1$, $0 < x < 1$

(b) $Y = -\log X$ and X has pdf

$$f_X(x) = \frac{(n+m+1)!}{n!\,m!}\, x^n(1-x)^m, \quad 0 < x < 1, \quad m, n \text{ positive integers}$$

(c) $Y = e^X$ and X has pdf

$$f_X(x) = \frac{1}{\sigma^2}\, xe^{-(x/\sigma)^2/2}, \quad 0 < x < \infty, \quad \sigma^2 \text{ a positive constant}$$

2.3 Suppose X has the geometric pmf $f_X(x) = \frac{1}{3}\left(\frac{2}{3}\right)^x$, $x = 0, 1, 2, \ldots$. Determine the probability distribution of $Y = X/(X+1)$. Note that here both X and Y are discrete random variables. To specify the probability distribution of Y, specify its pmf.

2.4 Let λ be a fixed positive constant, and define the function $f(x)$ by $f(x) = \frac{1}{2}\lambda e^{-\lambda x}$ if $x \geq 0$ and $f(x) = \frac{1}{2}\lambda e^{\lambda x}$ if $x < 0$.

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

(b) If X is a random variable with pdf given by $f(x)$, find $P(X < t)$ for all t. Evaluate all integrals.

(c) Find $P(|X| < t)$ for all t. Evaluate all integrals.

2.5 Use Theorem 2.1.8 to find the pdf of Y in Example 2.1.2. Show that the same answer is obtained by differentiating the cdf given in (2.1.6).

2.6 In each of the following find the pdf of Y and show that the pdf integrates to 1.

(a) $f_X(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$; $Y = |X|^3$

(b) $f_X(x) = \frac{3}{8}(x+1)^2$, $-1 < x < 1$; $Y = 1 - X^2$

(c) $f_X(x) = \frac{3}{8}(x+1)^2$, $-1 < x < 1$; $Y = 1 - X^2$ if $X \leq 0$ and $Y = 1 - X$ if $X > 0$

2.7 Let X have pdf $f_X(x) = \frac{2}{9}(x+1)$, $-1 \leq x \leq 2$.

(a) Find the pdf of $Y = X^2$. Note that Theorem 2.1.8 is not directly applicable in this problem.

(b) Show that Theorem 2.1.8 remains valid if the sets $A_0, A_1, \ldots, A_k$ *contain* $\mathcal{X}$, and apply the extension to solve part (a) using $A_0 = \emptyset$, $A_1 = (-2, 0)$, and $A_2 = (0, 2)$.

2.8 In each of the following show that the given function is a cdf and find $F_X^{-1}(y)$.

(a) $F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x} & \text{if } x \geq 0 \end{cases}$

(b) $F_X(x) = \begin{cases} e^x/2 & \text{if } x < 0 \\ 1/2 & \text{if } 0 \le x < 1 \\ 1 - (e^{1-x}/2) & \text{if } 1 \le x \end{cases}$

(c) $F_X(x) = \begin{cases} e^x/4 & \text{if } x < 0 \\ 1 - (e^{-x}/4) & \text{if } x \ge 0 \end{cases}$

Note that, in part (c), $F_X(x)$ is discontinuous but (2.1.13) is still the appropriate definition of $F_X^{-1}(y)$.

2.9 If the random variable X has pdf

$$f(x) = \begin{cases} \frac{x-1}{2} & 1 < x < 3 \\ 0 & \text{otherwise,} \end{cases}$$

find a monotone function $u(x)$ such that the random variable $Y = u(X)$ has a uniform$(0, 1)$ distribution.

2.10 In Theorem 2.1.10 the probability integral transform was proved, relating the uniform cdf to any continuous cdf. In this exercise we investigate the relationship between discrete random variables and uniform random variables. Let X be a discrete random variable with cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$.

(a) Prove that Y is stochastically greater than a uniform$(0, 1)$; that is, if $U \sim$ uniform $(0, 1)$, then

$$P(Y > y) \ge P(U > y) = 1 - y, \quad \text{for all } y,\ 0 < y < 1,$$
$$P(Y > y) > P(U > y) = 1 - y, \quad \text{for some } y,\ 0 < y < 1.$$

(Recall that *stochastically greater* was defined in Exercise 1.49.)

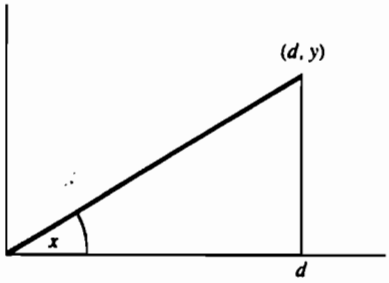
(b) Equivalently, show that the cdf of Y satisfies $F_Y(y) \le y$ for all $0 < y < 1$ and $F_Y(y) < y$ for some $0 < y < 1$. (*Hint:* Let x_0 be a jump point of F_X, and define $y_0 = F_X(x_0)$. Show that $P(Y \le y_0) = y_0$. Now establish the inequality by considering $y = y_0 + \epsilon$. Pictures of the cdfs will help.)

2.11 Let X have the standard normal pdf, $f_X(x) = (1/\sqrt{2\pi})e^{-x^2/2}$.

(a) Find EX^2 directly, and then by using the pdf of $Y = X^2$ from Example 2.1.7 and calculating EY.

(b) Find the pdf of $Y = |X|$, and find its mean and variance.

2.12 A random right triangle can be constructed in the following manner. Let X be a random angle whose distribution is uniform on $(0, \pi/2)$. For each X, construct a triangle as pictured below. Here, Y = height of the random triangle. For a fixed constant d, find the distribution of Y and EY.



2.13 Consider a sequence of independent coin flips, each of which has probability p of being heads. Define a random variable X as the length of the run (of either heads or tails) started by the first trial. (For example, $X = 3$ if either TTTH or HHHT is observed.) Find the distribution of X, and find $\mathrm{E}\,X$.

2.14 (a) Let X be a continuous, nonnegative random variable [$f(x) = 0$ for $x < 0$]. Show that

$$\mathrm{E}\,X = \int_0^\infty [1 - F_X(x)]\,dx,$$

where $F_X(x)$ is the cdf of X.

(b) Let X be a discrete random variable whose range is the nonnegative integers. Show that

$$\mathrm{E}\,X = \sum_{k=0}^{\infty} (1 - F_X(k)),$$

where $F_X(k) = P(X \le k)$. Compare this with part (a).

2.15 Betteley (1977) provides an interesting addition law for expectations. Let X and Y be any two random variables and define

$$X \wedge Y = \min(X, Y) \quad \text{and} \quad X \vee Y = \max(X, Y).$$

Analogous to the probability law $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, show that

$$\mathrm{E}(X \vee Y) = \mathrm{E}\,X + \mathrm{E}\,Y - \mathrm{E}(X \wedge Y).$$

(*Hint*: Establish that $X + Y = (X \vee Y) + (X \wedge Y)$.)

2.16 Use the result of Exercise 2.14 to find the mean duration of certain telephone calls, where we assume that the duration, T, of a particular call can be described probabilistically by $P(T > t) = ae^{-\lambda t} + (1-a)e^{-\mu t}$, where a, λ, and μ are constants, $0 < a < 1$, $\lambda > 0$, $\mu > 0$.

2.17 A *median* of a distribution is a value m such that $P(X \le m) \ge \frac{1}{2}$ and $P(X \ge m) \ge \frac{1}{2}$. (If X is continuous, m satisfies $\int_{-\infty}^{m} f(x)\,dx = \int_{m}^{\infty} f(x)\,dx = \frac{1}{2}$.) Find the median of the following distributions.

(a) $f(x) = 3x^2, \quad 0 < x < 1$ (b) $f(x) = \dfrac{1}{\pi(1+x^2)}, \quad -\infty < x < \infty$

2.18 Show that if X is a continuous random variable, then

$$\min_a \mathrm{E}\,|X - a| = \mathrm{E}\,|X - m|,$$

where m is the median of X (see Exercise 2.17).

2.19 Prove that

$$\frac{d}{da}\mathrm{E}(X-a)^2 = 0 \Leftrightarrow \mathrm{E}\,X = a$$

by differentiating the integral. Verify, using calculus, that $a = \mathrm{E}\,X$ is indeed a minimum. List the assumptions about F_X and f_X that are needed.

2.20 A couple decides to continue to have children until a daughter is born. What is the expected number of children of this couple? (*Hint:* See Example 1.5.4.)

2.21 Prove the "two-way" rule for expectations, equation (2.2.5), which says $\mathrm{E}\,g(X) = \mathrm{E}\,Y$, where $Y = g(X)$. Assume that $g(x)$ is a monotone function.

2.22 Let X have the pdf

$$f(x) = \frac{4}{\beta^3\sqrt{\pi}} x^2 e^{-x^2/\beta^2}, \quad 0 < x < \infty, \quad \beta > 0.$$

(a) Verify that $f(x)$ is a pdf. (b) Find $\mathrm{E}\,X$ and $\mathrm{Var}\,X$.

2.23 Let X have the pdf

$$f(x) = \frac{1}{2}(1+x), \quad -1 < x < 1.$$

(a) Find the pdf of $Y = X^2$. (b) Find $\mathrm{E}\,Y$ and $\mathrm{Var}\,Y$.

2.24 Compute $\mathrm{E}\,X$ and $\mathrm{Var}\,X$ for each of the following probability distributions.

(a) $f_X(x) = ax^{a-1}$, $0 < x < 1$, $a > 0$
(b) $f_X(x) = \frac{1}{n}$, $x = 1, 2, \ldots, n$, $n > 0$ an integer
(c) $f_X(x) = \frac{3}{2}(x-1)^2$, $0 < x < 2$

2.25 Suppose the pdf $f_X(x)$ of a random variable X is an *even function*. ($f_X(x)$ is an *even function* if $f_X(x) = f_X(-x)$ for every x.) Show that

(a) X and $-X$ are identically distributed.
(b) $M_X(t)$ is symmetric about 0.

2.26 Let $f(x)$ be a pdf and let a be a number such that, for all $\epsilon > 0$, $f(a+\epsilon) = f(a-\epsilon)$. Such a pdf is said to be *symmetric* about the point a.

(a) Give three examples of symmetric pdfs.
(b) Show that if $X \sim f(x)$, symmetric, then the median of X (see Exercise 2.17) is the number a.
(c) Show that if $X \sim f(x)$, symmetric, and $\mathrm{E}\,X$ exists, then $\mathrm{E}\,X = a$.
(d) Show that $f(x) = e^{-x}$, $x \geq 0$, is not a symmetric pdf.
(e) Show that for the pdf in part (d), the median is less than the mean.

2.27 Let $f(x)$ be a pdf, and let a be a number such that if $a \geq x \geq y$, then $f(a) \geq f(x) \geq f(y)$, and if $a \leq x \leq y$, then $f(a) \geq f(x) \geq f(y)$. Such a pdf is called *unimodal* with a *mode* equal to a.

(a) Give an example of a unimodal pdf for which the mode is unique.
(b) Give an example of a unimodal pdf for which the mode is not unique.
(c) Show that if $f(x)$ is both symmetric (see Exercise 2.26) and unimodal, then the point of symmetry is a mode.
(d) Consider the pdf $f(x) = e^{-x}$, $x \geq 0$. Show that this pdf is unimodal. What is its mode?

2.28 Let μ_n denote the nth central moment of a random variable X. Two quantities of interest, in addition to the mean and variance, are

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{3/2}} \quad \text{and} \quad \alpha_4 = \frac{\mu_4}{\mu_2^2}.$$

The value α_3 is called the *skewness* and α_4 is called the *kurtosis*. The skewness measures the lack of symmetry in the pdf (see Exercise 2.26). The kurtosis, although harder to interpret, measures the peakedness or flatness of the pdf.

(a) Show that if a pdf is symmetric about a point a, then $\alpha_3 = 0$.
(b) Calculate α_3 for $f(x) = e^{-x}$, $x \geq 0$, a pdf that is *skewed to the right.*
(c) Calculate α_4 for each of the following pdfs and comment on the peakedness of each.

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

$$f(x) = \frac{1}{2}, \quad -1 < x < 1$$

$$f(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < \infty$$

Ruppert (1987) uses *influence functions* (see Miscellanea 10.6.4) to explore further the meaning of kurtosis, and Groeneveld (1991) uses them to explore skewness; see also Balanda and MacGillivray (1988) for more on the interpretation of α_4.

2.29 To calculate moments of discrete distributions, it is often easier to work with the *factorial moments* (see Miscellanea 2.6.2).

(a) Calculate the factorial moment $\mathrm{E}[X(X-1)]$ for the binomial and Poisson distributions.
(b) Use the results of part (a) to calculate the variances of the binomial and Poisson distributions.
(c) A particularly nasty discrete distribution is the *beta-binomial*, with pmf

$$P(y = y) = a(y+a) \frac{\binom{n}{y}\binom{a+b-1}{a}}{\binom{n+a+b-1}{y+a}},$$

where n, a, and b are integers, and $y = 0, 1, 2, \ldots, n$. Use factorial moments to calculate the variance of the beta-binomial. (See Exercise 4.34 for another approach to this calculation.)

2.30 Find the moment generating function corresponding to

(a) $f(x) = \frac{1}{c}, \quad 0 < x < c.$

(b) $f(x) = \frac{2x}{c^2}, \quad 0 < x < c.$

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

(d) $P(X = x) = \binom{r+x-1}{x} p^r (1-p)^x, \quad x = 0, 1, \ldots, \quad 0 < p < 1, r > 0$ an integer.

2.31 Does a distribution exist for which $M_X(t) = t/(1-t), |t| < 1$? If yes, find it. If no, prove it.

2.32 Let $M_X(t)$ be the moment generating function of X, and define $S(t) = \log(M_X(t))$. Show that

$$\frac{d}{dt} S(t)\Big|_{t=0} = \mathrm{E}\, X \quad \text{and} \quad \frac{d^2}{dt^2} S(t)\Big|_{t=0} = \mathrm{Var}\, X.$$

2.33 In each of the following cases verify the expression given for the moment generating function, and in each case use the mgf to calculate $\mathrm{E}\, X$ and $\mathrm{Var}\, X$.

(a) $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $M_X(t) = e^{\lambda(e^t-1)}$, $x = 0, 1, \ldots$; $\lambda > 0$
(b) $P(X = x) = p(1-p)^x$, $M_X(t) = \frac{p}{1-(1-p)e^t}$, $x = 0, 1, \ldots$; $0 < p < 1$
(c) $f_X(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}$, $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$, $-\infty < x < \infty$; $-\infty < \mu < \infty$, $\sigma > 0$

2.34 A distribution cannot be uniquely determined by a finite collection of moments, as this example from Romano and Siegel (1986) shows. Let X have the normal distribution, that is, X has pdf

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

Define a discrete random variable Y by

$$P\left(Y = \sqrt{3}\right) = P\left(Y = -\sqrt{3}\right) = \frac{1}{6}, \qquad P(Y = 0) = \frac{2}{3}.$$

Show that

$$\mathrm{E}\, X^r = \mathrm{E}\, Y^r \quad \text{for } r = 1, 2, 3, 4, 5.$$

(Romano and Siegel point out that for any finite n there exists a discrete, and hence nonnormal, random variable whose first n moments are equal to those of X.)

2.35 Fill in the gaps in Example 2.3.10.

(a) Show that if $X_1 \sim f_1(x)$, then

$$\mathrm{E}\, X_1^r = e^{r^2/2}, \quad r = 0, 1, \ldots.$$

So $f_1(x)$ has all of its moments, and all of the moments are finite.

(b) Now show that

$$\int_0^\infty x^r f_1(x) \sin(2\pi \log x)\, dx = 0$$

for all positive integers r, so $\mathrm{E}X_1^r = \mathrm{E}\, X_2^r$ for all r. (Romano and Siegel 1986 discuss an extreme version of this example, where an entire class of distinct pdfs have the same moments. Also, Berg 1988 has shown that this moment behavior can arise with simpler transforms of the normal distribution such as X^3.)

2.36 The *lognormal distribution*, on which Example 2.3.10 is based, has an interesting property. If we have the pdf

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

then Exercise 2.35 shows that all moments exist and are finite. However, this distribution does not have a moment generating function, that is,

$$M_X(t) = \int_0^\infty \frac{e^{tx}}{\sqrt{2\pi}x} e^{-(\log x)^2/2}\, dx$$

does not exist. Prove this.

2.37 Referring to the situation described in Miscellanea 2.6.3:

(a) Plot the pdfs f_1 and f_2 to illustrate their difference.
(b) Plot the cumulant generating functions K_1 and K_2 to illustrate their similarity.

(c) Calculate the moment generating functions of the pdfs f_1 and f_2. Are they similar or different?

(d) How do the pdfs f_1 and f_2 relate to the pdfs described in Example 2.3.10?

2.38 Let X have the negative binomial distribution with pmf

$$f(x) = \binom{r+x-1}{x} p^r(1-p)^x, \quad x = 0, 1, 2, \ldots,$$

where $0 < p < 1$ and $r > 0$ is an integer.

(a) Calculate the mgf of X.

(b) Define a new random variable by $Y = 2pX$. Show that, as $p \downarrow 0$, the mgf of Y converges to that of a chi squared random variable with $2r$ degrees of freedom by showing that

$$\lim_{p \to 0} M_Y(t) = \left(\frac{1}{1-2t}\right)^r, \qquad |t| < \frac{1}{2}.$$

2.39 In each of the following cases calculate the indicated derivatives, justifying all operations.

(a) $\frac{d}{dx} \int_0^x e^{-\lambda t}\, dt$ (b) $\frac{d}{d\lambda} \int_0^\infty e^{-\lambda t}\, dt$

(c) $\frac{d}{dt} \int_t^1 \frac{1}{x^2}\, dx$ (d) $\frac{d}{dt} \int_1^\infty \frac{1}{(x-t)^2}\, dx$

2.40 Prove

$$\sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k} = (n-x)\binom{n}{x} \int_0^{1-p} t^{n-x-1}(1-t)^x\, dt.$$

(*Hint:* Integrate by parts or differentiate both sides with respect to p.)

2.6 Miscellanea

2.6.1 Uniqueness of Moment Sequences

A distribution is not necessarily determined by its moments. But if $\sum_{r=1}^\infty \mu^r r^k/k!$ has a positive radius of convergence, where $X \sim F_X$ and $\mathrm{E}\, X^r = \mu'_r$, then the moment sequence is unique, and hence the distribution is uniquely determined (Billingsley 1995, Section 30). Convergence of this sum also implies that the moment-generating function exists in an interval, and hence the moment-generating function determines the distribution

A sufficient condition for the moment sequence to be unique is *Carleman's Condition* (Chung 1974). If $X \sim F_X$ and we denote $\mathrm{E}\, X^r = \mu'_r$, then the moment sequence is unique if

$$\sum_{r=1}^{\infty} \frac{1}{(\mu'_{2r})^{1/(2r)}} = +\infty.$$

This condition is, in general, not easy to verify.

Feller (1971) has a very complete development of Laplace transforms, of which mgfs are a special case. In particular, Feller shows (similar to Billingsley) that whenever

$$M_X(t) = \sum_{r=0}^{\infty} \frac{(-1)^r \mu_r' t^r}{r!}$$

converges on an interval $-t_0 \leq t < t_0$, $t_0 > 0$, the distribution F_X is uniquely determined. Thus, when the mgf exists, the moment sequence determines the distribution F_X uniquely.

It should be clear that using the mgf to determine the distribution is a difficult task. A better method is through the use of *characteristic functions*, which are explained below. Although characteristic functions simplify the characterization of a distribution, they necessitate understanding complex analysis. You win some and you lose some.

2.6.2 Other Generating Functions

In addition to the moment generating function, there are a number of other generating functions available. In most cases, the characteristic function is the most useful of these. Except for rare circumstances, the other generating functions are less useful, but there are situations where they can ease calculations.

Cumulant generating function For a random variable X, the cumulant generating function is the function $\log[M_X(t)]$. This function can be used to generate the *cumulants* of X, which are defined (rather circuitously) as the coefficients in the Taylor series of the cumulant generating function (see Exercise 2.32).

Factorial moment generating function The factorial moment-generating function of X is defined as $\mathrm{E}t^X$, if the expectation exists. The name comes from the fact that this function satisfies

$$\frac{d^r}{dt^r} \mathrm{E}\, t^X \bigg|_{t=1} = \mathrm{E}\{X(X-1)\cdots(X-r+1)\},$$

where the right-hand side is a *factorial moment*. If X is a discrete random variable, then we can write

$$\mathrm{E}\, t^X = \sum_x t^x P(X = x),$$

and the factorial moment generating function is called the *probability-generating function*, since the coefficients of the power series give the probabilities. That is, to obtain the probability that $X = k$, calculate

$$\frac{1}{k!} \frac{d^k}{dt^k} \mathrm{E}\, t^X \bigg|_{t=1} = P(X = k).$$

Characteristic function Perhaps the most useful of all of these types of functions is the characteristic function. The characteristic function of X is defined by

$$\phi_X(t) = \mathrm{E}\, e^{itX},$$

where i is the complex number $\sqrt{-1}$, so the above expectation requires complex integration. The characteristic function does much more than the mgf does. When the moments of F_X exist, ϕ_X can be used to generate them, much like an mgf. The characteristic function always exists and it completely determines the distribution. That is, every cdf has a unique characteristic function. So we can state a theorem like Theorem 2.3.11, for example, but without qualification.

Theorem 2.6.1 (Convergence of characteristic functions) *Suppose X_k, $k = 1, 2, \ldots$, is a sequence of random variables, each with characteristic function $\phi_{X_k}(t)$. Furthermore, suppose that*

$$\lim_{k\to\infty} \phi_{X_k}(t) = \phi_X(t), \text{ for all } t \text{ in a neighborhood of } 0,$$

and $\phi_X(t)$ is a characteristic function. Then, for all X where $F_X(x)$ is continuous,

$$\lim_{k\to\infty} F_{X_k}(x) = F_X(x).$$

A full treatment of generating functions is given by Feller (1968). Characteristic functions can be found in almost any advanced probability text; see Billingsley (1995) or Resnick (1999).

2.6.3 Does the Moment Generating Function Characterize a Distribution?

In an article with the above title, McCullagh (1994) looks at a pair of densities similar to those in Example 2.3.10 but having mgfs

$$f_1 = n(0,1) \quad \text{and} \quad f_2 = f_1(x)\left[1 + \frac{1}{2}\sin(2\pi x)\right]$$

with cumulant generating functions

$$K_1(t) = t^2/2 \quad \text{and} \quad K_2(t) = K_1(t) + \log\left[1 + \frac{1}{2}e^{-2\pi^2}\sin(2\pi t)\right].$$

He notes that although the densities are visibly dissimilar, the cgfs are virtually identical, with maximum difference less than 1.34×10^{-9} over the entire range (less than the size of one pixel). So the answer to the question posed in the title is "*yes* for mathematical purposes but a resounding *no* for numerical purposes." In constrast, Waller (1995) illustrates that although the mgfs fail to numerically distinguish the distributions, the *characteristic functions* do a fine job. (Waller *et al.* 1995 and Luceño 1997 further investigate the usefulness of the characteristic function in numerically obtaining the cdfs.) See Exercise 2.37 for details.