Chapter 1

Introduction

1.1 Preliminaries

Definition (Differential equation)
A *differential equation* (DE) is an equation involving a function and its derivatives.

Differential equations are called *partial differential equations* (PDE) or *ordinary differential equations* (ODE) according to whether or not they contain partial derivatives. The *order* of a differential equation is the highest order derivative occurring. A *solution* (or *particular solution*) of a differential equation of order n consists of a function defined and n times differentiable on a domain D having the property that the functional equation obtained by substituting the function and its n derivatives into the differential equation holds for every point in D.

Example 1.1. An example of a differential equation of order 4, 2, and 1 is given respectively by

$$\left(\frac{dy}{dx}\right)^3 + \frac{d^4y}{dx^4} + y = 2\sin(x) + \cos^3(x),$$
$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0,$$
$$yy' = 1. \qquad *$$

Example 1.2. The function $y = \sin(x)$ is a solution of

$$\left(\frac{dy}{dx}\right)^3 + \frac{d^4y}{dx^4} + y = 2\sin(x) + \cos^3(x)$$

on domain $\mathbb{R}$; the function $z = e^x \cos(y)$ is a solution of

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = 0$$

on domain $\mathbb{R}^2$; the function $y = 2\sqrt{x}$ is a solution of

$$yy' = 2$$

on domain $(0, \infty)$. ∗

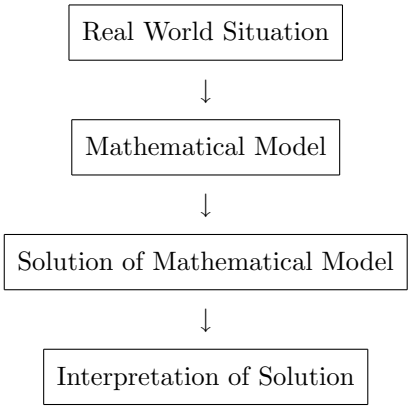
Although it is possible for a DE to have a unique solution, e.g., $y = 0$ is the solution to $(y')^2 + y^2 = 0$, or no solution at all, e.g., $(y')^2 + y^2 = -1$ has no solution, most DE's have infinitely many solutions.

Example 1.3. The function $y = \sqrt{4x + C}$ on domain $(-C/4, \infty)$ is a solution of $yy' = 2$ for any constant C. ∗

Note that different solutions can have different domains. The set of all solutions to a DE is call its *general solution*.

1.2 Sample Application of Differential Equations

A typical application of differential equations proceeds along these lines:



Sometimes in attempting to solve a DE, we might perform an irreversible step. This might introduce extra solutions. If we can get a short list which contains all solutions, we can then test out each one and throw out the invalid ones. The ultimate test is this: **does it satisfy the equation?**

Here is a sample application of differential equations.

Example 1.4. The half-life of radium is 1600 years, i.e., it takes 1600 years for half of any quantity to decay. If a sample initially contains 50 g, how long will it be until it contains 45 g? ∗

Solution. Let $x(t)$ be the amount of radium present at time t in years. The rate at which the sample decays is proportional to the size of the sample at that time. Therefore we know that $dx/dt = kx$. This differential equation is our mathematical model. Using techniques we will study in this course (see §3.2, Chapter 3), we will discover that the general solution of this equation is given by the equation $x = Ae^{kt}$, for some constant A. We are told that $x = 50$ when $t = 0$ and so substituting gives $A = 50$. Thus $x = 50e^{kt}$. Solving for t gives $t = \ln(x/50)/k$. With $x(1600) = 25$, we have $25 = 50e^{1600k}$. Therefore,

$$1600k = \ln\left(\frac{1}{2}\right) = -\ln(2),$$

giving us $k = -\ln(2)/1600$. When $x = 45$, we have

$$t = \frac{\ln(x/50)}{k} = \frac{\ln(45/50)}{-\ln(2)/1600} = -1600 \cdot \frac{\ln(8/10)}{\ln(2)} = 1600 \cdot \frac{\ln(10/8)}{\ln(2)}$$
$$\approx 1600 \cdot \frac{0.105}{0.693} \approx 1600 \times 0.152 \approx 243.2.$$

Therefore, it will be approximately 243.2 years until the sample contains 45 g of radium. ♢

Additional conditions required of the solution ($x(0) = 50$ in the above example) are called *boundary conditions* and a differential equation together with boundary conditions is called a *boundary-value problem* (BVP). Boundary conditions come in many forms. For example, $y(6) = y(22)$; $y'(7) = 3y(0)$; $y(9) = 5$ are all examples of boundary conditions. Boundary-value problems, like the one in the example, where the boundary condition consists of specifying the value of the solution at some point are also called *initial-value problems* (IVP).

Example 1.5. An analogy from algebra is the equation

$$y = \sqrt{y} + 2. \tag{1.1}$$

To solve for y, we proceed as

$$\begin{aligned} y - 2 &= \sqrt{y}, \\ (y-2)^2 &= y, \quad \text{(irreversible step)} \\ y^2 - 4y + 4 &= y, \\ y^2 - 5y + 4 &= 0, \\ (y-1)(y-4) &= 0. \end{aligned}$$

Thus, the set $y \in \{1, 4\}$ contains all the solutions. We quickly see that $y = 4$ satisfies Equation (1.1) because

$$4 = \sqrt{4} + 2 \Longrightarrow 4 = 2 + 2 \Longrightarrow 4 = 4,$$

while $y = 1$ does not because

$$1 = \sqrt{1} + 2 \Longrightarrow 1 = 3.$$

So we accept $y = 4$ and reject $y = 1$. $*$

Chapter 2

First Order Ordinary Differential Equations

The complexity of solving DE's increases with the order. We begin with first order DE's.

2.1 Separable Equations

A first order ODE has the form $F(x, y, y') = 0$. In theory, at least, the methods of algebra can be used to write it in the form* $y' = G(x, y)$. If $G(x, y)$ can be factored to give $G(x, y) = M(x)\,N(y)$,then the equation is called *separable*. To solve the separable equation $y' = M(x)\,N(y)$, we rewrite it in the form $f(y)y' = g(x)$. Integrating both sides gives

$$\int f(y)y'\,dx = \int g(x)\,dx,$$
$$\int f(y)\,dy = \int f(y)\frac{dy}{dx}\,dx.$$

Example 2.1. Solve $2xy + 6x + \left(x^2 - 4\right)y' = 0$. *

*We use the notation $dy/dx = G(x, y)$ and $dy = G(x, y)\,dx$ interchangeably.

Solution. Rearranging, we have

$$\begin{aligned}
\left(x^2-4\right)y' &= -2xy-6x,\\
&= -2xy-6x,\\
\frac{y'}{y+3} &= -\frac{2x}{x^2-4}, \quad x\neq\pm 2\\
\ln(|y+3|) &= -\ln\left(\left|x^2-4\right|\right)+C,\\
\ln(|y+3|)+\ln\left(\left|x^2-4\right|\right) &= C,
\end{aligned}$$

where C is an arbitrary constant. Then

$$\begin{aligned}
\left|(y+3)\left(x^2-4\right)\right| &= A,\\
(y+3)\left(x^2-4\right) &= A,\\
y+3 &= \frac{A}{x^2-4},
\end{aligned}$$

where A is a constant (equal to $\pm e^C$) and $x\neq\pm 2$. Also $y=-3$ is a solution (corresponding to $A=0$) and the domain for that solution is $\mathbb{R}$. ◊

Example 2.2. Solve the IVP $\sin(x)\,dx+y\,dy=0$, where $y(0)=1$. ∗

Solution. Note: $\sin(x)\,dx+y\,dy=0$ is an alternate notation meaning the same as $\sin(x)+y\,dy/dx=0$.

We have

$$\begin{aligned}
y\,dy &= -\sin(x)\,dx,\\
\int y\,dy &= \int -\sin(x)\,dx,\\
\frac{y^2}{2} &= \cos(x)+C_1,\\
y &= \sqrt{2\cos(x)+C_2},
\end{aligned}$$

where C_1 is an arbitrary constant and $C_2=2C_1$. Considering $y(0)=1$, we have

$$1=\sqrt{2+C_2}\Longrightarrow 1=2+C_2\Longrightarrow C_2=-1.$$

Therefore, $y=\sqrt{2\cos(x)-1}$ on the domain $(-\pi/3,\pi/3)$, since we need $\cos(x)\geq 1/2$ and $\cos(\pm\pi/3)=1/2$.

An alternate method to solving the problem is

$$\begin{aligned} y\,dy &= -\sin(x)\,dx, \\ \int_1^y y\,dy &= \int_0^x -\sin(x)\,dx, \\ \frac{y^2}{2} - \frac{1^2}{2} &= \cos(x) - \cos(0), \\ \frac{y^2}{2} - \frac{1}{2} &= \cos(x) - 1, \\ \frac{y^2}{2} &= \cos(x) - \frac{1}{2}, \\ y &= \sqrt{2\cos(x) - 1}, \end{aligned}$$

giving us the same result as with the first method. ◊

Example 2.3. Solve $y^4y' + y' + x^2 + 1 = 0$. ∗

Solution. We have

$$\begin{aligned} \left(y^4 + 1\right) y' &= -x^2 - 1, \\ \frac{y^5}{5} + y &= -\frac{x^3}{3} - x + C, \end{aligned}$$

where C is an arbitrary constant. This is an implicit solution which we cannot easily solve explicitly for y in terms of x. ◊

2.2 Exact Differential Equations

Using algebra, any first order equation can be written in the form $F(x, y)\,dx + G(x, y)\,dy = 0$ for some functions $F(x, y)$, $G(x, y)$.

Definition
An expression of the form $F(x, y)\,dx + G(x, y)\,dy$ is called a *(first-order) differential form*. A differentical form $F(x, y)\,dx + G(x, y)\,dy$ is called *exact* if there exists a function $g(x, y)$ such that $dg = F\,dx + G\,dy$.

If $\omega = F\,dx + G\,dy$ is an exact differential form, then $\omega = 0$ is called an *exact differential equation*. Its solution is $g = C$, where $\omega = dg$.

Recall the following useful theorem from `MATB42`:

Theorem 2.4
If F and G are functions that are continuously differentiable throughout a simply connected region, then $F\,dx + G\,dy$ is exact if and only if $\partial G/\partial x = \partial F/\partial y$.

Proof. Proof is given in `MATB42`. □

Example 2.5. Consider $\left(3x^2y^2 + x^2\right)dx + \left(2x^3y + y^2\right)dy = 0$. Let

$$\omega = \underbrace{\left(3x^2y^2 + x^2\right)}_{F}dx + \underbrace{\left(2x^3y + y^2\right)}_{G}dy$$

Then note that

$$\frac{\partial G}{\partial x} = 6x^2y = \frac{\partial F}{\partial y}.$$

By Theorem 2.4, $\omega = dg$ for some g. To find g, we know that

$$\frac{\partial g}{\partial x} = 3x^2y^2 + x^2, \tag{2.1a}$$

$$\frac{\partial g}{\partial y} = 2x^3y + y^2. \tag{2.1b}$$

Integrating Equation (2.1a) with respect to x gives us

$$g = x^3y^2 + \frac{x^3}{3} + h(y). \tag{2.2}$$

So differentiating that with respect to y gives us

$$\begin{aligned}
\overbrace{\frac{\partial g}{\partial y}}^{\text{Eq. (2.1b)}} &= 2x^3y + \frac{dh}{dy}, \\
2x^3y + y^2 &= 2x^3y + \frac{dh}{dy}, \\
\frac{dh}{dy} &= y^2, \\
h(y) &= \frac{y^3}{3} + C
\end{aligned}$$

for some arbitrary constant C. Therefore, Equation (2.2) becomes

$$g = x^3y^2 + \frac{x^3}{3} + \frac{y^3}{3} + C.$$

Note that according to our differential equation, we have

$$d\left(x^3y^2 + \frac{x^3}{3} + \frac{y^3}{3} + C\right) = 0 \text{ which implies } x^3y^2 + \frac{x^3}{3} + \frac{y^3}{3} + C = C'$$

for some arbitrary constant C'. Letting $D = C' - C$, which is still an arbitrary constant, the solution is

$$x^3y^2 + \frac{x^3}{3} + \frac{y^3}{3} = D. \qquad *$$

Example 2.6. Solve $\left(3x^2 + 2xy^2\right) dx + \left(2x^2y\right) dy = 0$, where $y(2) = -3$. *

Solution. We have

$$\int \left(3x^2 + 2xy^2\right) dx = x^3 + x^2y^2 + C$$

for some arbitrary constant C. Since C is arbitrary, we equivalently have $x^3 + x^2y^2 = C$. With the initial condition in mind, we have

$$8 + 4 \cdot 9 = C \Longrightarrow C = 44.$$

Therefore, $x^3 + x^2y^2 = 44$ and it follows that

$$y = \frac{\pm\sqrt{44 - x^3}}{x^2}.$$

But with the restriction that $y(2) = -3$, the only solution is

$$y = -\frac{\sqrt{44 - x^3}}{x^2}$$

on the domain $\left(-\sqrt[3]{44}, \sqrt[3]{44}\right) \setminus \{0\}$. ◊

Let $\omega = F\,dx + G\,dy$. Let $y = s(x)$ be the solution of the DE $\omega = 0$, i.e., $F + Gs'(x) = 0$. Let $y_0 = s(x_0)$ and let γ be the piece of the graph of $y = s(x)$ from (x_0, y_0) to (x, y). Figure 2.1 shows this idea. Since $y = s(x)$ is a solution to $\omega = 0$, we must have $\omega = 0$ along γ. Therefore, $\int_\gamma \omega = 0$. This can be seen

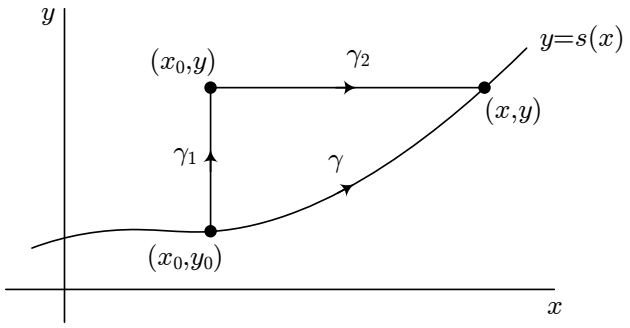


Figure 2.1: The graph of $y = s(x)$ with γ connecting (x_0, y_0) to (x, y).

by parameterizing γ by $\gamma(x) = (x, s(x))$, thereby giving us

$$\int_\gamma \omega = \int_{x_0}^{x} F\,dx + Gs'(x)\,dx = \int_{x_0}^{x} 0\,dx = 0.$$

This much holds for any ω.

Now suppose that ω is exact. Then the integral is independent of the path. Therefore

$$\begin{aligned} 0 = \int_\gamma \omega &= \int_{\gamma_1} F\,dx + G\,dy + \int_{\gamma_2} F\,dx + G\,dy \\ &= \int_{y_0}^{y} G(x_0, y)\,dy + \int_{x_0}^{x} F(x, y)\,dx. \end{aligned}$$

We can now solve Example 2.6 with this new method.

Solution (Alternate solution to Example 2.6). We simply have

$$\begin{aligned} 0 &= \int_{-3}^{4} 2 \cdot 2^2 y\,dy + \int_2^x \left(3x^2 + 2xy^2\right) dx \\ &= 4y^2 - 4\,(-3)^2 + x^3 + x^2y^2 - 2^3 - 2^2y^2 \\ &= 4y^2 - 36 + x^3 + x^2y^2 - 8 - 4y^2, \end{aligned}$$

finally giving us $x^3 + x^2y^2 = 44$, which agrees with our previous answer. $\lozenge$

Remark. Separable equations are actually a special case of exact equations, that is,

$$f(y)y' = g(x) \Longrightarrow -g(x)\,dx + f(y)\,dy = 0 \Longrightarrow \frac{\partial}{\partial x} f(y) = 0 = \frac{\partial}{\partial y}\left(-g(x)\right).$$

So the equation is exact. ◊

2.3 Integrating Factors

Consider the equation $\omega = 0$. Even if ω is not exact, there may be a function $I(x, y)$ such that $I\omega$ is exact. So $\omega = 0$ can be solved by multiplying both sides by I. The function I is called an *integrating factor* for the equation $\omega = 0$.

Example 2.7. Solve $y/x^2 + 1 + y'/x = 0$. ∗

Solution. We have

$$\left(\frac{y}{x^2} + 1\right) dx + \frac{1}{x}\, dy = 0.$$

We see that

$$\left[\frac{\partial}{\partial x}\left(\frac{1}{x}\right) = -\frac{1}{x^2}\right] \neq \left[\frac{1}{x^2} = \frac{\partial}{\partial y}\left(\frac{y}{x^2} + 1\right)\right].$$

So the equation is not exact. Multiplying by x^2 gives us

$$\begin{aligned} \left(y + x^2\right) dx + x\, dy &= 0, \\ d\left(xy + \frac{x^3}{3}\right) &= 0, \\ xy + \frac{x^3}{3} &= C \end{aligned}$$

for some arbitrary constant C. Solving for y finally gives us

$$y = \frac{C}{x} - \frac{x^3}{3}.$$ ◊

There is, in general, no algorithm for finding integrating factors. But the following may suggest where to look. It is important to be able to recognize common exact forms:

$$\begin{aligned} x\, dy + y\, dx &= d(xy)\,, \\ \frac{x\, dy - y\, dx}{x^2} &= d\left(\frac{y}{x}\right), \\ \frac{x\, dx + y\, dy}{x^2 + y^2} &= d\left(\frac{\ln\left(x^2 + y^2\right)}{2}\right), \\ \frac{x\, dy - d\, dx}{x^2 + y^2} &= d\left(\tan^{-1}\left(\frac{y}{x}\right)\right), \\ x^{a-1}y^{b-1}\left(ay\, dx + bx\, dy\right) &= d\left(x^a y^b\right). \end{aligned}$$

Example 2.8. Solve $\left(x^2y^2+y\right)dx+\left(2x^3y-x\right)dy=0.$ $*$

Solution. Expanding, we have

$$x^2y^2\,dx+2x^3y\,dy+y\,dx-x\,dy=0.$$

Here, $a=1$ and $b=2$. Thus, we wish to use

$$d\big(xy^2\big)=y^2\,dx+2xy\,dy.$$

This suggests dividing the original equation by x^2 which gives

$$y^2\,dx+2xy\,dy+\frac{y\,dx-x\,dy}{x^2}=0.$$

Therefore,

$$xy^2+\frac{y}{x}=C,\quad x\neq 0,$$

where C is an arbitrary constant. Additionally, $y=0$ on the domain $\mathbb{R}$ is a solution to the original equation. $\lozenge$

Example 2.9. Solve $y\,dx-x\,dy-\left(x^2+y^2\right)dx=0.$ $*$

Solution. We have

$$\frac{y\,dx-x\,dy}{x^2+y^2}-dx=0,$$

unless $x=0$ and $y=0$. Now, it follows that

$$\begin{aligned}
-\tan^{-1}\Big(\frac{y}{x}\Big)-x&=C,\\
\tan^{-1}\Big(\frac{y}{x}\Big)&=-C-x,\\
\tan^{-1}\Big(\frac{y}{x}\Big)&=D-x,\quad (D=-C)\\
\frac{y}{x}&=\tan(D-x)\,,\\
y&=x\tan(D-x)\,,
\end{aligned}$$

where C is an arbitrary constant and the domain is

$$D-x\neq(2n+1)\,\frac{\pi}{2},\quad x\neq(2n+1)\,\frac{\pi}{2}$$

for any integer n. Also, since the derivation of the solution is based on the assumption that $x\neq 0$, it is unclear whether or not 0 should be in the domain, i.e., does $y=x\tan(D-x)$ satisfy the equation when $x=0$? We have $y-xy'-$

$\left(x^2+y^2\right)=0$. If $x=0$ and $y=x\tan(D-x)$, then $y=0$ and the equation is satisfied. Thus, 0 is in the domain. ◊

Proposition 2.10
Let $\omega = dg$. Then for any function $P:\mathbb{R}\to\mathbb{R}$, $P(g)$ is exact.

Proof. Let $Q=\int P(t)\,dy$. Then $d(Q(g)) = P(g)\,dg = P(g)\omega$. □

To make use of Proposition 2.10, we can group together some terms of ω to get an expression you recognize as having an integrating factor and multiply the equation by that. The equation will now look like $dg+h=0$. If we can find an integrating factor for h, it will not necessarily help, since multiplying by it might mess up the part that is already exact. But if we can find one of the form $P(g)$, then it will work.

Example 2.11. Solve $\left(x-yx^2\right)dy + y\,dx = 0$. ∗

Solution. Expanding, we have

$$\underbrace{y\,dx + x\,dy}_{d(xy)} - yx^2\,dy = 0.$$

Therefore, we can multiply teh equation by any function of xy without disturbing the exactness of its first two terms. Making the last term into a function of y alone will make it exact. So we multiply by $(xy)^{-2}$, giving us

$$\frac{y\,dx+x\,dy}{x^2y^2} - \frac{1}{y}\,dy = 0 \Longrightarrow -\frac{1}{xy} - \ln(|y|) = C,$$

where C is an arbitrary constant. Note that $y=0$ on the domain $\mathbb{R}$ is also a solution. ◊

Given

$$M\,dx + N\,dy = 0, \qquad (*)$$

we want to find I such that $IM\,dx + IN\,dy$ is exact. If so, then

$$\underbrace{\frac{\partial}{\partial x}(IN)}_{I_xN+IN_x} = \underbrace{\frac{\partial}{\partial y}(IM)}_{I_yM+IM_y}.$$

If we can find any particular solution $I(x, y)$ of the PDE

$$I_x N + I N_x = I_y M + I M_y, \qquad (**)$$

then we can use it as an integrating factor to find the general solution of $(*)$. Unfortunately, $(**)$ is usually even harder to solve than $(*)$, but as we shall see, there are times when it is easier.

Example 2.12. We could look for an I having only x's and no y's? For example, consider $I_y = 0$. Then

$$I_x N + I N_x = I M_y \text{ implies } \frac{I_x}{I} = \frac{M_y - N_x}{N}.$$

This works if $(M_y - N_x)/N$ happens to be a function of x alone. Then

$$I = e^{\int \frac{M_y - N_x}{N}\, dx}.$$

Similarly, we can also reverse the role of x and y. If $(N_x - M_y)/M$ happens to be a function of y alone, then

$$e^{\int \frac{N_x - M_y}{M}\, dy}$$

works. $*$

2.4 Linear First Order Equations

A first order linear equation $(n = 1)$ looks like

$$y' + P(x)y = Q(x).$$

An integrating factor can always be found by the following method. Consider

$$\begin{aligned} dy + P(x)y\, dx &= Q(x)\, dx, \\ \underbrace{(P(x)y - Q(x))}_{M(x,y)}\, dx + \underbrace{dy}_{N(x,y)} &= 0. \end{aligned}$$

We use the DE for the integrating factor $I(x, y)$. The equation $IM\, dx + IN\, dy$ is exact if

$$I_x N + I N_x = I_y M + I M_y.$$

In our case,

$$I_x + 0 = I_y\,(P(x)y - Q(x)) + IP(x). \qquad (*)$$

We need only one solution, so we look for one of the form $I(x)$, i.e., with $I_y = 0$. Then $(*)$ becomes

$$\frac{dI}{dx} = IP(x).$$

This is separable. So

$$\begin{aligned} \frac{dI}{I} &= P(x)\,dx, \\ \ln(|I|) &= \int P(x)\,dx + C, \\ |I| &= e^{\int P(x)\,dx}, \quad e^x > 0 \\ I &= e^{\int P(x)\,dx}. \end{aligned}$$

We conclude that $e^{\int P(x)\,dx}$ is an integrating factor for $y' + P(x)y = Q(x)$.

Example 2.13. Solve $y' - (1/x)\,y = x^3$, where $x > 0$. $*$

Solution. Here $P(x) = -1/x$. Then

$$I = e^{\int P(x)\,dx} = e^{-\int \frac{1}{x}\,dx} = e^{-\ln(|x|)dx} = \frac{1}{|x|} = \frac{1}{x},$$

where $x > 0$. Our differential equation is

$$\frac{x\,dy - y\,dx}{x} = x^3\,dx.$$

Multiplying by the integrating factor $1/x$ gives us

$$\frac{x\,dy - y\,dx}{x^2} = x^2\,dx.$$

Then

$$\begin{aligned} \frac{y}{x} &= \frac{x^3}{3} + C, \\ y &= \frac{x^3}{3} + Cx \end{aligned}$$

on the domain $(0, \infty)$, where C is an arbitrary constant ($x > 0$ is given). $\Diamond$

2 MATH

Probability (Pure math 6)

Prepared by

Kena Math. Department

Chapter 1 : Introduction To Probability

$$PV = RT = 62.36T$$

or $$P(5) = 62.36(100)$$

$$P = 1247.2 \text{ mm mercury}$$

That is, our model leads us to expect the pressure to be 1247.2 mm mercury. A model such as the Perfect Gas Law is said to be "deterministic." It is deterministic in the sense that it allows us to determine an exact value for the variable of interest under specified experimental conditions. The Perfect Gas Law does describe *some* real gases at moderate temperatures and pressures. Unfortunately, many real gases cannot be described by this or any other deterministic model, especially at extreme temperatures and pressures! Under these circumstances we must find another way to predict the behavior of the gas with some degree of certainty. This can be done with the aid of statistical methods.

What do we mean by statistical methods? These are methods by which decisions are made based on the analysis of data gathered in carefully designed experiments. Since experiments cannot be designed to account for every conceivable contingency, there is always some uncertainty in experimental science. Statistical methods are designed to *allow us to assess the degree of uncertainty present in our results*. These methods can be classed roughly into three categories: descriptive statistics, inferential statistics, and model building. By descriptive statistics we mean those techniques, both analytic and graphical, that allow us to describe or picture a data set. Inferential statistics concerns methods by which conclusions can be drawn about a large group of objects, based on observing only a portion of the objects in the larger group.

This idea leads to the following definition:

Definition: The overall group of objects about which conclusions are to be drawn is called the *population*. A subset or portion of the population that is actually obtained and that is used to draw conclusions about the population is called a *sample*.

Model building entails the development of prediction equations from experimental data. These equations are called statistical models; they are models that allow us to predict the behavior of a complex system and to assess our probability of error. These categories are not mutually exclusive. That is, methods developed to solve problems in one area often find application in another. We shall be concerned with all three areas in this text.

A statistician or user of statistics is always working in two worlds. The ideal world is at the population level and is theoretical in nature. It is the world that we would like to see. The world of reality is the sample world. This is the level at which we really operate. We hope that the characteristics of our sample reflect well the characteristics of the population. That is, we treat our sample as a microcosm that mirrors the population. This idea is illustrated in Fig. 1.1.

The mathematics on which statistical methods rest is called probability theory. For this reason, we begin the study of statistics by considering the basic concepts of probability.

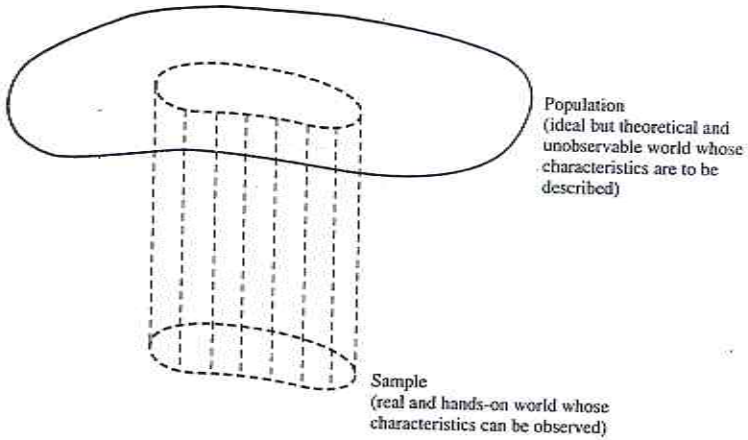


FIGURE 1.1
The sample is viewed as a miniature population. We hope that the behavior of the variable under study over the sample gives an accurate picture of its behavior in the population.

1.1 INTERPRETING PROBABILITIES

When asked, "Do you know anything about probability?" most people are quick to answer, "no!" Usually that is not the case at all. The ability to interpret probabilities is assumed in our culture. One hears the phrases "the probability of rain today is 95%" or "there is a 0% chance of rain today." It is assumed that the general public can interpret these values correctly. The interpretation of probabilities is summarized as follows:

Interpretation of Probabilities

1. Probabilities are numbers between 0 and 1, inclusive, that reflect the chances of a physical event occurring.
2. Probabilities near 1 indicate that the event is extremely likely to occur. They mean not that the event will occur, only that the event is considered to be a common occurrence.
3. Probabilities near zero indicate that the event is not very likely to occur. They do not mean that the event will fail to occur, only that the event is considered to be rare.
4. Probabilities near 1/2 indicate that the event is just as likely to occur as not.
5. Since numbers between 0 and 1 can be expressed as percentages between 0 and 100, probabilities are often expressed as percentages. This is particularly common in writings of a nontechnical nature.

These properties are guidelines for interpreting probabilities once they are available, but they do not indicate how to assign probabilities to events. Three

methods are widely used: the *personal* approach, the *relative frequency* approach, and the *classical approach*. These methods are illustrated in the following examples.

Example 1.1.1. An oil spill has occurred. An environmental scientist asks, "What is the probability that this spill can be contained before it causes widespread damage to nearby beaches?" Many factors come into play, among them the type of spill, the amount of oil spilled, the wind and water conditions during the clean-up operation, and the nearness of the beaches. These factors make this spill unique. The scientist is called upon to make a value judgment, that is, to assign a probability to the event based on informed *personal opinion*.

The main advantage of the personal approach is that it is always applicable. Anyone can have a personal opinion about anything. Its main disadvantage is, of course, that its accuracy depends on the accuracy of the information available and the ability of the scientist to assess that information correctly. --

Example 1.1.2. An electrical engineer is studying the peak demand at a power plant. It is observed that on 80 of the 100 days randomly selected for study from past records, the peak demand occurred between 6 and 7 p.m. It is natural to assume that the probability of this occurring on another day is at least *approximately*

$$\frac{80}{100} = .80$$

This figure is not simply a personal opinion. It is a figure based on repeated experimentation and observation. It is a *relative frequency*.

The relative frequency approach can be used whenever the experiment can be repeated many times and the results observed. In such cases, the probability of the occurrence of event A, denoted by $P[A]$, is approximated as follows:

Relative Frequency Approximation

$$P[A] \doteq \frac{f}{n} = \frac{\text{number of times event } A \text{ occurred}}{\text{number of times experiment was run}}$$

The disadvantage in this approach is that the experiment cannot be a one-shot situation; it must be repeatable. Remember that any probability obtained this way is an approximation. It is a value based on n trials. Further testing might result in a different approximate value. However, as the number of trials increases, the changes in the approximate values obtained tend to become slight. Thus for a large number of trials, the approximate probability obtained by using the relative frequency approach is usually quite accurate.

Example 1.1.3. What is the probability that a child born to a couple heterozygous for eye color (each with genes for both brown and blue eyes) will be brown-eyed? To answer this question, we note that since the child receives one gene from each parent, the possibilities for the child are (brown, blue), (blue, brown), (blue, blue) and

(brown, brown), where the first member of each pair represents the gene received from the father. Since each parent is just as likely to contribute a gene for brown eyes as for blue eyes, all four possibilities are equally likely. Since the gene for brown eyes is dominant, three of the four possibilities lead to a brown-eyed child. Hence the probability that the child will be brown-eyed is 3/4 = .75.

The above probability is not a personal opinion, nor is it based on repeated experimentation. In fact, we found this probability by the *classical* method. This method can be used *only* when it is reasonable to assume that the possible outcomes of the experiment are equally likely. In this case, the probability of the occurrence of event A is given by the following classical formula:

Classical Formula

$$P[A] = \frac{n(A)}{n(S)} = \frac{\text{number of ways } A \text{ can occur}}{\text{number of ways the experiment can proceed}}$$

One advantage to this method is that it does not require experimentation. Furthermore, if the outcomes are truly equally likely, then the probability assigned to event A is not an approximation. It is an accurate description of the frequency with which event A will occur.

1.2 SAMPLE SPACES AND EVENTS

To determine what is "probable" in an experiment, we first must determine what is "possible." That is, the first step in analyzing most experiments is to make a list of possibilities for the experiment. Such a list is called a *sample space*. We define this term as follows:

Definition 1.2.1 (Sample space and sample point). A sample space for an experiment is a set S with the property that each physical outcome of the experiment corresponds to exactly one element of S. An element of S is called a sample point.

When the number of possibilities is small, an appropriate sample space usually can be found without difficulty. For instance, we have seen that when a couple heterozygous for eye color parents a child, the possible genotypes for the child are given by

$$S = \{(\text{brown, blue}), (\text{blue, brown}), (\text{blue, blue}), (\text{brown, brown})\}$$

As the number of possibilities becomes larger, it is helpful to have a system for developing a sample space. One such system is the *tree diagram*. The next example illustrates the idea.

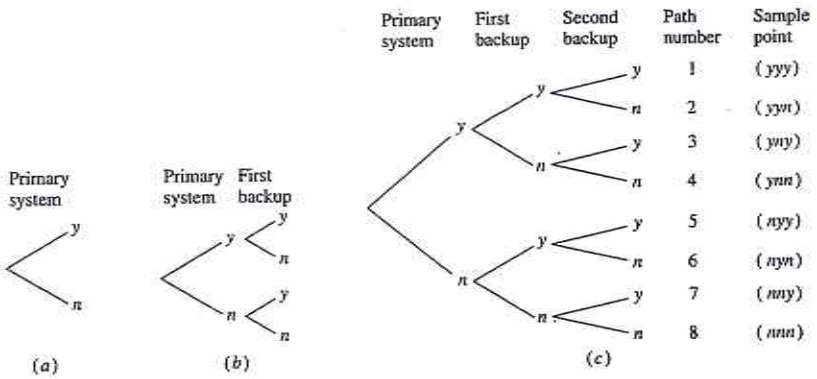


FIGURE 1.2
Constructing a tree diagram.

Example 1.2.1. During a space shot the primary computer system is backed up by two secondary systems. They operate independently of one another in that the failure of one has no effect on any of the others. We are interested in the readiness of these three systems at launch time. What is an appropriate sample space for this experiment?

Since we are primarily concerned with whether each system is operable at launch, we need only find a sample space that gives that information. To generate the sample space, we use a *tree*. The primary system is either operable (yes) or not operable (no) at the time of launch. This is indicated in the tree diagram of Fig. 1.2(a), where yes = y and no = n. Likewise the first backup system either is or is not operable. This is shown in Fig. 1.2(b). Finally, the second backup system either is or is not operable. The tree is completed as shown in Fig. 1.2(c). A sample space S for the experiment can be read from the tree by following each of the eight distinct paths through the tree. Thus

$$S = \{yyy, yyn, yny, ynn, nyy, nyn, nny, nnn\}$$

Once a suitable sample space has been found, elementary set theory can be used to describe physical occurrences associated with the experiment. This is done by considering what are called *events* in the mathematical sense.

Definition 1.2.2 (Event). Any subset A of a sample space is called an event. The empty set $\varnothing$ is called the *impossible* event; the subset S is called the *certain* event.

Example 1.2.2. Consider a space shot in which a primary computer system is backed up by two secondary systems. The sample space for this experiment is

$$S = \{yyy, yyn, yny, ynn, nyy, nyn, nny, nnn\}$$

where, for example, *yny* denotes the fact that the primary system and second backup are operable at launch, whereas the first backup is inoperable (see Example 1.2.1). Let

A: primary system is operable
B: first backup is operable
C: second backup is operable

The mathematical event corresponding to each of these physical events is found by listing the sample points that represent the occurrence of the event. Thus we write

$$A = \{yyy, yyn, yny, ynn\}$$
$$B = \{yyy, yyn, nyy, nyn\}$$
$$C = \{yyy, yny, nyy, nny\}$$

Other events can be described using these events as building blocks. For example, the event that "the primary system *or* the first backup is operable" is given by the set $A \cup B$, the union of set A with set B. Recall from elementary mathematics that the *union of A with B consists of all sample points that are in set A or set B or are in both.* Thus

$$A \cup B = \frac{\text{primary } or \text{ first}}{\text{backup is operable}} = \{yyy, yyn, yny, ynn, nyy, nyn\}$$

Note that the word "or" will denote set union. The event that "the primary system *and* the first backup is operable" is given by the set $A \cap B$, the intersection of set A with set B. *The intersection of two sets consists of all sample points that are in both sets.* That is, it is the set of points that they have in common. Here

$$A \cap B = \text{primary } and \text{ first backup operable} = \{yyy, yyn\}$$

Note that the word "and" will denote the set intersection. The event that "the primary system or the first backup is operable but the second backup is inoperable" is given by $(A \cap B) \cap C'$, where C' denotes the complement of set C. *The complement of a set consists of the sample points in the sample space that are not in the given set.* Thus

$$(A \cup B) \cap C' = \frac{\text{primary } or \text{ first backup operable}}{but \text{ second backup inoperable}} = \{yyn, ynn, nyn\}$$

Note that the word "but" is also translated as a set intersection; the word "not" translates as a set complement.

Let us pause briefly to consider a basic difference between the sample space

$$S_1 = \{(\text{brown, blue}), (\text{blue, brown}), (\text{blue, blue}), (\text{brown, brown})\}$$

of Example 1.1.3 and

$$S_2 = \{yyy, yyn, yny, ynn, nyy, nyn, nny, nnn\}$$

of Example 1.2.1. Since each parent is just as likely to contribute a gene for brown eyes as for blue eyes, the sample points of S_1 are equally likely. This allows us to use the classical method to find the probability that a child born to a couple heterozygous for eye color will be brown-eyed. If we denote this event by A, then we can conclude that

$$P[A] = P[\{(\text{brown, blue}), (\text{blue, brown}), (\text{brown, brown})\}]$$
$$= \frac{n(A)}{n(S)} = \frac{3}{4}$$

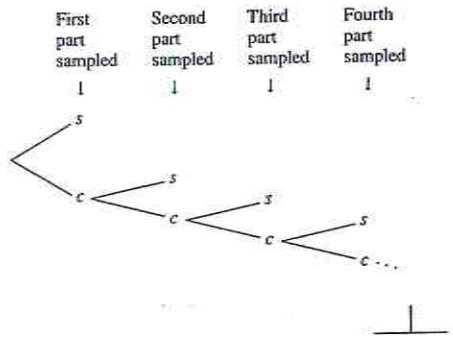


FIGURE 1.3
Sampling a production line for defective parts.

However, it is not correct to assume that the sample points of S_2 are equally likely. This would be true if and only if each of the three computer systems is just as likely to fail as to be operable at launch time. Our technology is much better than that! The primary question to be answered is "What is the probability that at least one system will be operable at the time of the launch?" That is, what is

$$P[\{yyy, yyn, yny, ynn, nyy, nyn, nny\}]?$$

As will be shown later, this question can be answered. However, since the sample points are not equally likely, it cannot be answered using the classical method.

Not all trees are symmetric as is that pictured in Fig. 1.2. In some settings, paths end at different stages of the game. Example 1.2.3 illustrates an experiment of this sort.

Example 1.2.3. Consider a production process that is known to produce defective parts at the rate of one per hundred. The process is monitored by testing randomly selected parts during the production process. Suppose that as soon as a defective part is found, the process will be stopped and all machine settings will be checked. We are interested in studying the number of parts that are tested in order to obtain the first defective part. In the tree of Fig. 1.3, c represents that the sampling continues and s represents that production is stopped. Notice that as soon as a defective item is found, the process ends and the path also ends. For this reason, some paths are much shorter than others. Notice also that theoretically this tree continues indefinitely. The sample space generated by the tree is

$$S = \{s, cs, ccs, cccs, ccccs, \ldots\}$$

Since defective parts occur with probability .01, it should be evident that the paths of this tree are not equally likely.

Mutually Exclusive Events

Occasionally interest centers on two or more events that cannot occur at the same time. That is, the occurrence of one event precludes the occurrence of the other. Such events are said to be *mutually exclusive.*

Example 1.2.4. Consider the sample space

$$S = \{yyy, yyn, yny, ynn, nyy, nyn, nny, nnn\}$$

of Example 1.2.1. The events

$$A_1\text{: primary system operable} = \{yyy, yyn, yny, ynn\}$$
$$A_2\text{: primary system inoperable} = \{nyy, nyn, nny, nnn\}$$

are mutually exclusive. It is impossible for the primary system to be both operable and inoperable at the same time. Mathematically, A_1 and A_2 have no sample points in common. That is, $A_1 \cap A_2 = \varnothing$.

Example 1.2.4 suggests the mathematical definition of the term "mutually exclusive events."

Definition 1.2.3 (Mutually exclusive events). Two events A_1 and A_2 are mutually exclusive if and only if $A_1 \cap A_2 = \varnothing$. Events $A_1, A_2, A_3, \ldots$ are mutually exclusive if and only if $A_i \cap A_j = \varnothing$ for $i \neq j$.

1.3 PERMUTATIONS AND COMBINATIONS

As indicated in Sec. 1.1, there are several ways to determine the probability of an event. When the physical description of the experiment leads us to believe that the possible outcomes are equally likely, then we can compute the probability of the occurrence of an event using the classical method. In this case the probability of an event A is given by

$$P[A] = \frac{n(A)}{n(S)}$$

Thus to compute a probability using the classical approach, you must be able to count two things: $n(A)$, the number of ways in which event A can occur, and $n(S)$, the number of ways in which the experiment can proceed. As the experiment becomes more complex, lists and trees become cumbersome. Alternative methods for counting must be found.

Two types of counting problems are common. The first involves *permutations* and the second, *combinations*. These terms are defined as follows:

(*e*) What is the name given to an event such as D?
(*f*) If at any given time each switch is just as likely to be on as off, what is the probability that no switch is on?

36. Two items are randomly selected one at a time from an assembly line and classed as to whether they are of superior quality (+), average quality (0), or inferior quality (−).
(*a*) Construct a tree for this two-stage experiment.
(*b*) List the elements of the sample space generated by the tree.
(*c*) List the sample points that constitute the events
A: the first item selected is of inferior quality
B: the quality of each of the items is the same
C: the quality of the first item exceeds that of the second
(*d*) Are the events A and B mutually exclusive? Are the events A and C mutually exclusive?
(*e*) Give a brief verbal description of these events:
$A' \cap B$ $A' \cap B'$
$A \cap B'$ $A \cap C' \cap B$
(*f*) It is known that 90% of the items produced are of average quality, 1% are of superior quality, and the rest are of inferior quality. It is argued that since the classification experiment can proceed in nine ways with only one of these resulting in two items of average quality, the probability of obtaining two such items is 1/9. Criticize this argument.

37. An experiment consists of selecting a digit from among the digits 0 to 9 in such a way that each digit has the same chance of being selected as any other. We name the digit selected A. These lines of code are then executed:

IF $A < 2$ THEN $B = 12$; ELSE $B = 17$;
IF $B = 12$ THEN $C = A - 1$; ELSE $C = 0$;

(*a*) Construct a tree to illustrate the ways in which values can be assigned to the variables A, B, and C.
(*b*) Find the sample space generated by the tree.
(*c*) Are the 10 possible outcomes for this experiment equally likely?
(*d*) Find the probability that A is an even number.
(*e*) Find the probability that C is negative.
(*f*) Find the probability that $C = 0$.
(*g*) Find the probability that $C \leq 1$.

38. Consider Exercise 16. If experimental runs are to be done in random order, how many different sequences are possible? (Set up only!) In experiments of this sort, runs are not usually done randomly. Rather, they are carefully designed so that the researcher has control of the order of experimentation. Can you think of some practical reasons for why this is necessary?

CHAPTER 2

SOME PROBABILITY LAWS

In Chap. 1 we considered how to interpret probabilities. In this chapter we consider some laws that govern their behavior. The laws that we shall present are those that will have a direct application to problem solving. These laws will be stated and illustrated numerically. Their derivations are not hard, and most of them are left as exercises.

2.1 AXIOMS OF PROBABILITY

You have probably seen the development of a mathematical system in your study of high school geometry. In developing any mathematical system, one begins by stating a few basic definitions and axioms that underlie the system. The definitions are the technical terms of the system; axioms are statements that are assumed to be true and therefore require no proof. Usually one starts with as few axioms as possible and then uses these axioms and the technical definitions to develop whatever theorems follow logically. Some technical terms such as sample space, sample point, event, and mutually exclusive events have already been introduced. One can develop a useful system of theorems pertaining to probability with the aid of these definitions and three axioms, called the axioms of probability.

Axioms of probability.
1. Let S denote a sample space for an experiment:
$$P[S] = 1$$
2. $P[A] \geq 0$ for every event A.
3. Let $A_1, A_2, A_3, \ldots$ be a finite or an infinite collection of mutually exclusive events. Then $P[A_1 \cup A_2 \cup A_3 \cdots] = P[A_1] + P[A_2] + P[A_3] + \cdots$.

Axiom 1 states a fact that most people regard as obvious; namely, the probability assigned to the certain event S is 1. Axiom 2 ensures that probabilities can never be negative. Axiom 3 guarantees that when one deals with mutually exclusive events, the probability that at least one of the events will occur can be found by adding the individual probabilities. An important consequence of this axiom is that it gives us the ability to find the probability of an event when the sample points in the same space for the experiment are not equally likely. Example 2.1.1 illustrates this point.

Example 2.1.1. The distribution of blood types in the United States is roughly 41% type A, 9% type B, 4% type AB, and 46% type O. An individual is brought into an emergency room and is to be blood-typed. What is the probability that the type will be A, B, or AB?

The sample space for this experiment is

$$S = \{A, B, AB, O\}$$

The sample points are not equally likely, so the classical approach to probability is not applicable. That is, we cannot say that since there are four blood types and three of them are A, B, or AB the probability of obtaining one of these types is ¾. Let A_1, A_2, and A_3 denote the events that the patient has type A, B, and AB blood, respectively. The events A_1, A_2, and A_3 are mutually exclusive because one cannot have two different blood types at the same time. We are looking for $P[A_1 \cup A_2 \cup A_3]$. By axiom 3,

$$\begin{aligned} P[A_1 \cup A_2 \cup A_3] &= P[A_1] + P[A_2] + P[A_3] \\ &= .41 + .09 + .04 \\ &= .54 \end{aligned}$$

An immediate consequence of these axioms is the fact that the probability assigned to the impossible event is 0, as you should suspect. The derivation of this result is outlined in Exercise 12.

Theorem 2.1.1. $P[\varnothing] = 0$.

Another consequence of the axioms is that the probability that an event will *not* occur is equal to 1 minus the probability that it will occur. For example, if the probability of a successful space shuttle mission is .99, then the probability that it will not be successful is $1 - .99 = .01$. This idea is stated in Theorem 2.1.2. Its derivation is outlined in Exercise 12.

Theorem 2.1.2. $P[A'] = 1 - P[A]$.

The General Addition Rule

We have seen how to handle questions concerning the probability of one or another event occurring if those events are mutually exclusive. We now develop a more general rule that will allow us to find the probability that at least one of two events

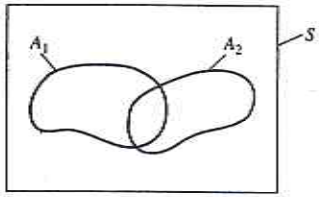


FIGURE 2.1
$A_1 \cap A_2 \neq \varnothing$.

will occur when the events are not necessarily mutually exclusive. This rule is suggested by considering the Venn diagram of Fig. 2.1. Assume that the shaded region in the diagram, $A_1 \cap A_2$, is not empty so that A_1 and A_2 are not mutually exclusive. If we claim that

$$P[A_1 \cup A_2] = P[A_1] + P[A_2]$$

we have committed an obvious error. Since $A_1 \cap A_2$ is contained in A_1 and $A_1 \cap A_2$ is contained in A_2, $P[A_1 \cap A_2]$ has been included twice in our calculation. To correct this error, we subtract $P[A_1 \cap A_2]$ from the right-hand side of the equation to obtain the general addition rule:

General addition rule

$$P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$$

This rule can be derived from the axioms of probability and the theorems that we have already developed. Its proof is outlined in Exercise 12. The key word that signals its use is the word "or."

Example 2.1.2. Components of a propulsion system can be arranged in series. However, this arrangement has a serious drawback; if one component fails, the system fails. This is obviously a risky arrangement for space travel! Consider a system in which the main engine has a backup. These engines are designed to operate independently in that the success or failure of one has no effect on the other. The engine component is operable if one *or* the other of these two engines is operable. Such a system is said to have the engine component in parallel. Assume that each engine is 90% reliable. That is, each functions correctly with probability .9. As we shall show later, it is then reasonable to assume that both engines operate correctly with probability .81. Find the probability that the engine component is operable. Let A_1: the main engine is operable, and A_2: the backup engine is operable. We are given that $P[A_1] = P[A_2] = .9$ and that $P[A_1 \cap A_2] = .81$. We want to find $P[A_1 \cup A_2]$. By the addition rule

$$\begin{aligned} P[A_1 \cup A_2] &= P[A_1] + P[A_2] - P[A_1 \cap A_2] \\ &= .9 + .9 - .81 = .99 \end{aligned}$$

The addition rule links the operations of union and intersection. If $P[A_1 \cap A_2]$ is known, the addition rule can be used to find $P[A_1 \cup A_2]$. Similarly, if $P[A_1 \cup A_2]$

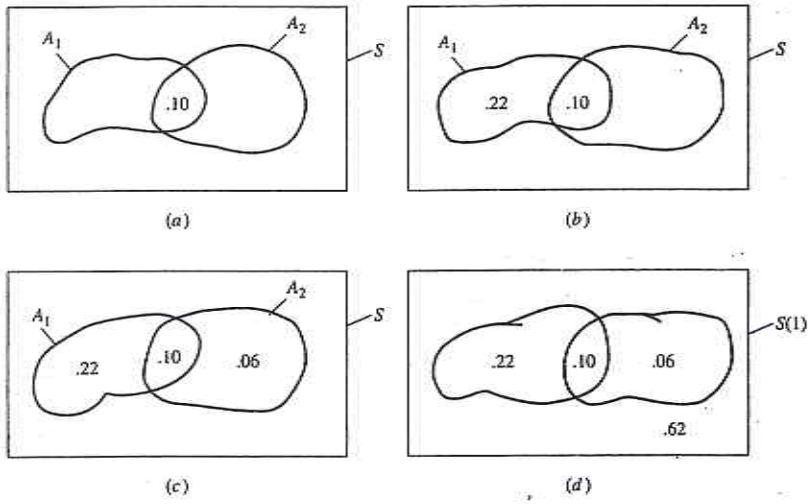


FIGURE 2.2
(a) $P[A_1 \cap A_2] = .10$; (b) $P[A_1 \cap A_2'] = .22$; (c) $P[A_1' \cap A_2] = .06$; (d) $P[A_1' \cap A_2'] = .62$.

is known, we can use the rule to find $P[A_1 \cap A_2]$. Venn diagrams are helpful when using this rule.

Example 2.1.3. A chemist analyzes seawater samples for two heavy metals: lead and mercury. Past experience indicates that 38% of the samples taken from near the mouth of a river on which numerous industrial plants are located contain toxic levels of lead or mercury: 32% contain toxic levels of lead and 16% contain toxic levels of mercury. What is the probability that a randomly selected sample will contain toxic levels of lead only? Let A_1 denote the event that the sample contains toxic levels of lead, and let A_2 denote that the sample contains toxic levels of mercury. We are given that $P[A_1] = .32$, $P[A_2] = .16$, and $P[A_1 \cup A_2] = .38$. By the addition rule

$$P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$$

or

$$.38 = .32 + .16 - P[A_1 \cap A_2]$$

Solving this equation, we obtain $P[A_1 \cap A_2] = .10$. This is indicated in Fig. 2.2(*a*). Since $P[A_1] = .32$ and $A_1 \cap A_2$ is contained in A_1, the probability associated with the shaded region in Fig. 2.2(*b*) is .22. Similarly, since $A_1 \cap A_2$ is contained in A_2, a probability of .06 is associated with the shaded region of Fig. 2.2(*c*). Finally, since $P[S] = 1$, the probability assigned to the shaded area in Fig. 2.2(*d*) is .62. We are asked to find the probability that the sample will contain only lead. That is, we want to find $P[A_1 \cap A_2']$. This probability, .22, can be read from Fig. 2.2(*b*).

Notice that if the percentages reported in problems such as these are based on population data, then the probabilities calculated by use of the general addition rule are exact. However, if the percentages reported are based on samples drawn from a

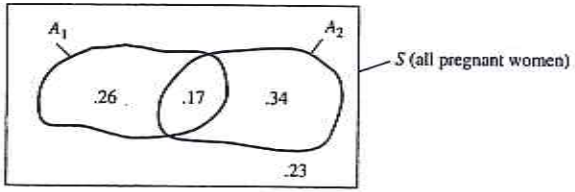


FIGURE 2.3
Partition of S.

larger population, then the probabilities computed are relative frequency probabilities. They are *approximations* to the true probability of the occurrence of the event in question. Since most percentages reported in the literature are based on samples, most of them are properly viewed as being relative frequency probabilities. We use the word "probability" with the understanding that the probabilities given and computed by using the theorems in this chapter are, in most cases, only approximations.

2.2 CONDITIONAL PROBABILITY

In this section we introduce the notion of conditional probability. The name itself is indicative of what is to be done. We wish to determine the probability that some event A_2 will occur, "conditional on" the assumption that some other event A_1 has occurred. The key words to look for in identifying a conditional question are "if" and "given that." We use the notation $P[A_2|A_1]$ to denote the conditional probability of event A_2 occurring given that event A_1 has occurred. A simple example will suggest the way to define this probability.

Example 2.2.1. In trying to determine the sex of a child a pregnancy test called "starch gel electrophoresis" is used. This test may reveal the presence of a protein zone called the pregnancy zone. This zone is present in 43% of all pregnant women. Furthermore, it is known that 51% of all children born are male. Seventeen percent of all children born are male and the pregnancy zone is present. The Venn diagram for these data is shown in Fig. 2.3. Let A_1 denote the event that the pregnancy zone is present, and A_2 that the child is male. We know that, for a randomly selected pregnant woman, $P[A_1] = .43$, $P[A_2] = .51$, $P[A_1 \cap A_2] = .17$. If asked, "What is the probability that the child is male?" the answer is .51. Suppose we are *given* the information that the pregnancy zone is present and asked, "What is the probability that the child is male?" We now have information that was not available originally. What effect, if any, does this new information have on our belief that the child is male? That is, what is $P[A_2|A_1]$? Once we know that the pregnancy zone is present, our sample space no longer includes all pregnant women; it consists only of the 43% with this characteristic. Of these, $.17/.43 \doteq .395$ have male children. Logic implies that

$$P[\text{male}|\text{zone present}] = P[A_2|A_1] = .395$$

Receipt of the information that the pregnancy zone is present reduces from .51 to .395 the probability that the child is male.

To formalize the reasoning used in the previous example, note that $P[A_2|A_1]$ is found by forming a ratio whose denominator is $P[A_1]$, the probability that the *given* event will occur. The numerator is $P[A_1 \cap A_2]$, the probability that *both* the given event and the event in question will occur. That is, we define the conditional probability as follows:

Definition 2.2.1 (Conditional probability). Let A_1 and A_2 be events such that $P[A_1] \neq 0$. The conditional probability of A_2 given A_1, denoted by $P[A_2|A_1]$, is defined by

$$P[A_2|A_1] = \frac{P[A_1 \cap A_2]}{P[A_1]}$$

Sometimes receipt of the information that event A_1 has occurred has no effect on the probability assigned to event A_2. That is,

$$P[A_2|A_1] = P[A_2]$$

When this happens, A_1 and A_2 have a special relationship to one another. The nature of this relationship will be explored in the next section. In the meantime don't be surprised if you find that a particular conditional probability does not differ from the original probability assigned to the event!

2.3 INDEPENDENCE AND THE MULTIPLICATION RULE

We have used the word "independent" informally in several previous examples. Webster's dictionary defines independent objects as objects acting "irrespective of each other." Thus two events are independent if one may occur irrespective of the other. That is, the occurrence or nonoccurrence of one does not alter the likelihood of occurrence or nonoccurrence of the other. In some cases it is reasonable to assume that two events are independent from the physical description of the events themselves. For example, suppose that a couple heterozygous for eye color has two children. Since the eye color of a child is affected only by the genetic makeup of the parents and not by the eye color of the other child, it is reasonable to assume that the events A_1: the first child has brown eyes, and A_2: the second child has brown eyes, are independent. However, in most instances the issue is not clear-cut. In these cases we need a mathematical definition of the term to determine without a doubt whether two events are, in fact, independent.

To see how to characterize independence, let us consider a simple experiment that consists of rolling a single fair die once and then tossing a fair coin once. Let the first member of each ordered pair denote the number appearing on the die and the second, the face showing on the coin (H = heads, T = tails). A sample space for this experiment is

$$S = \{(1, H), (1, T), (2, H), (2, T), (3, H), (3, T), (4, H), (4, T), (5, H), (5, T), (6, H), (6, T)\}$$

Since the die and the coin are considered to be fair, these 12 outcomes are equally likely. Consider these events:

A: the die shows one or two

B: the coin shows heads

$A \cap B$: the die shows one or two and the coin shows heads

Since knowing the result of the die roll gives us no additional information on how the coin will land, it is reasonable to assume that the events A and B are independent. Using classical probability, we easily see that

$$P[A] = P[\{(1, H), (1, T), (2, H), (2, T)\}] = 4/12 = 1/3$$
$$P[B] = P[\{(1, H), (2, H), (3, H), (4, H), (5, H), (6, H)\}] = 6/12 = 1/2$$
$$P[A \cap B] = P[\{(1, H), (2, H)\}] = 2/12 = 1/6$$

More importantly, it is easy to see that for these physically independent events

$$P[A \cap B] = P[A] \cdot P[B]$$

Consider now an experiment that consists of drawing two coins in succession from a box containing a nickel (N), a dime (D), and a quarter (Q). The first coin is not replaced before the second is drawn. A sample space for this experiment is

$$S = \{(N, D), (N, Q), (D, N), (D, Q), (Q, N), (Q, D)\}$$

These outcomes are equally likely. Consider these events:

A: the first coin is a dime

B: the second coin is a dime

Since we do not replace the first coin before the second draw, it is evident that if event A occurs, event B cannot occur. That is, knowledge that event A has occurred does give us information on whether or not event B will occur! These events are not independent. Using classical probability, we easily see that

$$P[A] = P[\{(D, N), (D, Q)\}] = 2/6$$
$$P[B] = P[\{(N, D), (Q, D)\}] = 2/6$$
$$P[A \cap B] = P[\varnothing] = 0$$

More importantly, it is easy to see that for these events that are not independent

$$P[A \cap B] \neq P[A]P[B]$$

Thus we have noticed that when A and B are clearly independent, $P[A \cap B] = P[A]P[B]$; when they are clearly dependent, $P[A \cap B] \neq P[A]P[B]$. This is not coincidental. It is natural to use this mathematical characterization as our technical definition of the term "independent events."

Definition 2.3.1 (Independent events). Events A_1 and A_2 are independent if and only if

$$P[A_1 \cap A_2] = P[A_1]P[A_2]$$

This definition is useful in two ways. If exact probabilities are available, then it serves as a test for independence. However, since most probabilities encountered in scientific studies are approximations, it is most useful as a way to find the probability that two events will occur when the events are clearly independent. Example 2.3.1 illustrates its use as a test for independence.

Example 2.3.1. Consider the experiment of drawing a card from a well-shuffled deck of 52 cards. Let

A_1: a spade is drawn

A_2: an honor (10, J, Q, K, A) is drawn

Classical probability is used to see that $P[A_1] = 13/52$ and $P[A_2] = 20/52$. The probability that a spade and an honor, $P[A_1 \cap A_2]$, is drawn is 5/52. Notice that these probabilities are exact. They are not approximations based on observations of card draws. Are the events A_1 and A_2 independent? To decide, note that

$$P[A_1]P[A_2] = (13/52)(20/52) = 5/52$$

and

$$P[A_1 \cap A_2] = 5/52$$

Since $P[A_1 \cap A_2] = P[A_1]P[A_2]$, we can conclude that these events are independent.

In Chap. 15 a test for independence will be developed that can be used when working with real data rather than with classical probabilities. Its derivation is based on the definition of independent events just discussed.

Example 2.3.2 illustrates the use of Definition 2.3.1 in finding the probability that two events will occur simultaneously when the events are clearly independent.

Example 2.3.2. In Example 1.1.3, we found that the probability that a couple heterozygous for eye color will parent a brown-eyed child is 3/4 for each child. Genetic studies indicate that the eye color of one child is independent of that of the other. Thus if the couple has two children, then the probability that both will be brown-eyed is

$$P\left[\begin{matrix}\text{first}\\ \text{brown}\end{matrix} \text{ and } \begin{matrix}\text{second}\\ \text{brown}\end{matrix}\right] = P\left[\begin{matrix}\text{first}\\ \text{brown}\end{matrix}\right]P\left[\begin{matrix}\text{second}\\ \text{brown}\end{matrix}\right]$$
$$= \frac{3}{4} \cdot \frac{3}{4}$$
$$= \frac{9}{16}$$

Definition 2.3.1 defines independence for *any* events A_1 and A_2. If at least one of the events A_1 or A_2 occurs with *nonzero* probability, then an appealing

characterization of independence can be obtained. To see how this is done, assume that $P[A_1] \neq 0$. By Definition 2.3.1, A_1 and A_2 are independent if and only if

$$P[A_1 \cap A_2] = P[A_1]P[A_2]$$

Dividing by $P[A_1]$, we can conclude that A_1 and A_2 are independent if and only if

$$\frac{P[A_1 \cap A_2]}{P[A_1]} = P[A_2|A_1] = P[A_2]$$

A similar argument holds if $P[A_2] \neq 0$. We have thus derived the result given in Theorem 2.3.1.

Theorem 2.3.1. Let A_1 and A_2 be events such that at least one of $P[A_1]$ or $P[A_2]$ is nonzero. A_1 and A_2 are independent if and only if

$$P[A_2|A_1] = P[A_2] \quad \text{if } P[A_1] \neq 0 \quad \text{and}$$
$$P[A_1|A_2] = P[A_1] \quad \text{if } P[A_2] \neq 0$$

Since most events of real interest do occur with nonzero probability, Theorem 2.3.1 is used as a test for independence. To understand the logic behind the theorem, let us reconsider the data of Example 2.3.1.

Example 2.3.3. Consider the events A_1, a spade is drawn, and A_2, an honor is drawn. We know that $P[A_1] = 13/52$, $P[A_2] = 20/52$, and $P[A_1 \cap A_2] = 5/52$. Suppose we are asked, "What is the probability that a randomly selected card is an honor?" Our answer is 20/52. Suppose we are now told that the card is a spade and are asked, "What is the probability that the card is an honor?" That is, "What is $P[A_2|A_1]$?" If A_1 and A_2 are independent, the new information is irrelevant and our answer should not change. That is, $P[A_2|A_1] = P[A_2]$. Otherwise our answer should change, and $P[A_2|A_1] \neq P[A_2]$. In this setting, is $P[A_2|A_1] = P[A_2]$? To answer this question, note that

$$P[A_2|A_1] = \frac{P[A_1 \cap A_2]}{P[A_1]} = \frac{5/52}{13/52} = 5/13$$

and

$$P[A_2] = 20/52 = 5/13$$

Since these probabilities are the same, we conclude via Theorem 2.3.1 that A_1 and A_2 are independent.

Occasionally we must deal with more than two events. Again, the question arises, "When are these events considered independent?" Definition 2.3.2 answers this question by extending our previous definition to include more than two events.

Definition 2.3.2. Let $C = \{A_i: i = 1, 2, \ldots, n\}$ be a finite collection of events. These events are independent if and only if, given any subcollection $A_{(1)}, A_{(2)}, \ldots, A_{(m)}$ of elements of C,

$$P[A_{(1)} \cap A_{(2)} \cap \cdots \cap A_{(m)}] = P[A_{(1)}]P[A_{(2)}] \cdots P[A_{(m)}]$$

Although this definition can be used to test a collection of events for independence, its main purpose is to provide a way to find the probability that a series of events that are assumed to be independent will occur. To illustrate, we reconsider a problem encountered in Chap. 1 (Example 1.2.1).

Example 2.3.4. During a space shot, the primary computer system is backed up by two secondary systems. They operate independently of one another, and each is 90% reliable. What is the probability that all three systems will be operable at the time of the launch? Let

A_1: the main system is operable

A_2: the first backup is operable

A_3: the second backup is operable

We are given that $P[A_1] = P[A_2] = P[A_3] = .9$. We want $P[A_1 \cap A_2 \cap A_3]$. Since these events are assumed to be independent,

$$\begin{aligned} P[A_1 \cap A_2 \cap A_3] &= P[A_1]P[A_2]P[A_3] \\ &= (.9)(.9)(.9) \\ &= .729 \end{aligned}$$

Definition 2.3.2 must be used with care. In particular, one must be certain that it is reasonable to assume that events are independent before it is applied to compute the probability that a series of events will occur. The danger of erroneously assumed independence is illustrated in Example 2.3.5.

Example 2.3.5. An Atomic Energy Commission Study, WASH 1400, reported the probability of a nuclear accident such as that which occurred at Three Mile Island in March 1978 to be one in 10 million. Yet the accident did occur. According to Mark Stephens, "The methodology of WASH 1400 made use of event trees—sequences of actions that would be necessary for accidents to take place. These event trees did not assume any interrelation between events—that they might be caused by the same error in judgment or as part of the same mistaken action. The statisticians who assigned probabilities in the writing of WASH 1400 said, for example, that there was a one-in-a-thousand risk of one of the auxiliary feed-water control valves—the twelves—being closed. And if there is a one-in-a-thousand chance of one valve being closed, the chances of both valves being closed is one-thousandth of that, or a million to one. But both of the twelves were closed by the same man on March 26—and one had never been closed without the other." The events A_1: the first valve is closed, and A_2: the second valve is closed were not independent. However, they were treated as such when calculating the probability of an accident. This, among other things, led to an underestimate of the accident potential (from *Three Mile Island* by Mark Stephens, Random House, 1980).

The Multiplication Rule

There is one further point to be made before we conclude this section. We can find $P[A_1 \cap A_2]$ if the events are assumed to be independent. Furthermore, if the proper information is given, the general addition rule can be used to find this probability.

Is there any other way to find the probability of the simultaneous occurrence of two events if the events are not independent? The answer is yes, and the method is easy to derive. We know that

$$P[A_2|A_1] = \frac{P[A_1 \cap A_2]}{P[A_1]} \qquad P[A_1] \neq 0$$

regardless of whether the events are independent. Multiplying each side of this equation by $P[A_1]$, we obtain the following formula, called the *multiplication rule:*

Multiplication rule

$$P[A_1 \cap A_2] = P[A_2|A_1]P[A_1]$$

The use of this rule is illustrated in Example 2.3.6.

Example 2.3.6. Recent research indicates that approximately 49% of all infections involve anaerobic bacteria. Furthermore, 70% of all anaerobic infections are polymicrobic; that is, they involve more than one anaerobe. What is the probability that a given infection involves anaerobic bacteria *and* is polymicrobic? Let A_1 denote the event that the infection is anaerobic, and A_2 that it is polymicrobic. We are given that $P[A_1] = .49$ and that $P[A_2|A_1] = .70$. We want to find $P[A_1 \cap A_2]$. By the multiplication rule,

$$\begin{aligned} P[A_1 \cap A_2] &= P[A_2|A_1]P[A_1] \\ &= (.70)(.49) \\ &= .343 \end{aligned}$$

2.4 BAYES' THEOREM

The topic of this section is the theorem formulated by the Reverend Thomas Bayes (1761). It deals with conditional probability. Bayes' theorem is used to find $P[A|B]$ when the available information is not immediately compatible with that required to apply the definition of conditional probability directly.

Example 2.4.1 is a typical problem calling for the use of Bayes' theorem. You will find applying Bayes' rule quite natural without having seen a formal statement of the theorem!

Example 2.4.1. Assume that 40% of all interstate highway accidents involve excessive speed on the part of at least one of the drivers (event E) and that 30% involve alcohol use by at least one driver (event A). If alcohol is involved there is a 60% chance that excessive speed is also involved; otherwise, this probability is only 10%. An accident involves speeding. What is the probability that alcohol is involved? We are given these probabilities:

$$P[E] = .40 \quad P[A] = .30 \quad P[E|A] = .60$$
$$P[E'] = .60 \quad P[A'] = .70 \quad P[E|A'] = .10$$

We are being asked to find $P[A|E]$. Since this is a conditional question, it is natural to turn to the definition of conditional probability for a solution. In this case,

$$P[A|E] = \frac{P[E \cap A]}{P[E]}$$

Unfortunately, neither of the probabilities needed for the solution is immediately available. However, each can be obtained easily. By the multiplication rule,

$$P[E \cap A] = P[E|A]P[A]$$

Note that if excessive speed was involved, alcohol use either was or was not also involved. Hence event E can be subdivided into two mutually exclusive events as follows:

$$E = (E \cap A) \cup (E \cap A')$$

Thus
$$P[E] = P[E \cap A] + P[E \cap A']$$

An expression has already been found for the first probability on the right; the multiplication rule can be applied to the second probability to see that

$$P[E \cap A'] = P[E|A']P[A']$$

Substitution now yields

$$P[A|E] = \frac{P[E \cap A]}{P[E]}$$
$$= \frac{P[E|A]P[A]}{P[E|A]P[A] + P[E|A']P[A']}$$

Note the pattern in this solution. In the numerator the conditional expression is the reverse of that in the original question; in the denominator, the conditional expressions run through all of the alternatives to the event in question, in this case A and A'. The numerical solution can now be obtained by substitution as follows:

$$P[A|E] = \frac{P[E|A]P[A]}{P[E|A]P[A] + P[E|A']P[A']}$$
$$= \frac{(.60)(.30)}{(.60)(.30) + (.10)(.70)}$$
$$= .72$$

If excessive speed was involved in an accident, there is a 72% chance that alcohol was also involved.

In the previous example, there were two mutually exclusive events, A and A', whose union is S. Bayes' theorem can also be applied when S is subdivided into more than two mutually exclusive events. We state the theorem in this more general setting.

Theorem 2.4.1 (Bayes' theorem). Let $A_1, A_2, A_3, \ldots, A_n$ be a collection of mutually exclusive events whose union is S. Let B be an event such that $P[B] \neq 0$. Then for any of the events $A_j, j = 1, 2, 3, \ldots, n$,

$$P[A_j|B] = \frac{P[B|A_j]P[A_j]}{\sum_{i=1}^{n} P[B|A_i]P[A_i]}$$

To see that Bayes' theorem could have been used directly to answer the question posed in Example 2.4.1, note that events A and A' are mutually exclusive events whose union is S and that event E occurs with nonzero probability. Hence we can make the following identifications:

$$A_1 = A \qquad A_2 = A' \qquad B = E$$

By applying Bayes' theorem directly we obtain

$$P[A_1|B] = \frac{P[B|A_1]P[A_1]}{P[B|A_1]P[A_1] + P[B|A_2]P[A_2]}$$

or
$$P[A|E] = \frac{P[E|A]P[A]}{P[E|A]P[A] + P[E|A']P[A']}$$

A quick comparison will show that this is the same as the solution derived in Example 2.4.1 using the multiplication rule.

The next example illustrates the use of Bayes' theorem in a setting in which the sample space is subdivided into four mutually exclusive events rather than two.

Example 2.4.2. The blood type distribution in the United States is type A, 41%; type B, 9%; type AB, 4%; and type O, 46%. It is estimated that during World War II, 4% of inductees with type O blood were typed as having type A; 88% of those with type A were correctly typed; 4% with type B blood were typed as A; and 10% with type AB were typed as A. A soldier was wounded and brought to surgery. He was typed as having type A blood. What is the probability that this is his true blood type? Let

A_1: he has type A blood
A_2: he has type B blood
A_3: he has type AB blood
A_4: he has type O blood
B: he is typed as type A

Note that the events A_1, A_2, A_3, A_4 are mutually exclusive, and their union is S because each individual can have only one blood type and all possible blood types have been listed. We are being asked to find $P[A_1|B]$. We are given that

$$P[A_1] = .41 \quad P[B|A_1] = .88$$
$$P[A_2] = .09 \quad P[B|A_2] = .04$$
$$P[A_3] = .04 \quad P[B|A_3] = .10$$
$$P[A_4] = .46 \quad P[B|A_4] = .04$$

Substitution into the expression given by Bayes' theorem yields

$$P[A_1|B] = \frac{(.88)(.41)}{(.88)(.41) + (.04)(.09) + (.10)(.04) + (.04)(.46)}$$
$$\doteq .93$$

If a person was typed as having type A blood, there was approximately a 93% chance that his true type was in fact type A.

Chapter 3 : Discrete Distributions

of polyethylene glycol, a fusion-promoting agent. It is known that the probability that such a cell will fuse is 1/2. Let Y denote the number of cells exposed to obtain the first fusion. The variable Y is random; a priori it can assume any value in the set $\{1, 2, 3, \ldots\}$. Recall from your study of calculus that a set such as this that consists of an infinite collection of isolated points is called a *countably infinite set.*

Example 3.1.3. In Example 1.1.2 we considered the variable T, the time at which the peak demand for electricity occurs per day. This variable is random, since its value is affected by such chance factors as time of the year, humidity, and temperature. It can conceivably assume any value in the 24-hour time span from 12 midnight one day to 12 midnight the next day.

It is easy to distinguish a discrete random variable from one that is not discrete. Just ask the question, "What are the possible values for the variable?" If the answer is a finite set or a countably infinite set, then the random variable is discrete; otherwise it is not. This idea leads to the following definition:

Definition 3.1.1 (Discrete random variable). A random variable is discrete if it can assume at most a finite or a countably infinite number of possible values.

The random variable X, the number of brown-eyed children in a two-child family, is discrete. Its set of possible values is the finite set $\{0, 1, 2\}$. The set $\{1, 2, 3, \ldots\}$ of possible values for Y, the number of cells exposed to obtain the first fusion of Example 3.1.2, is countably infinite. Thus Y is also a discrete random variable. The random variable T, the time of the peak demand for electricity at a power plant, is different from the others. Time is measured continuously, and T can conceivably assume any value in the interval $[0, 24)$, where 0 denotes 12 midnight one day and 24 denotes 12 midnight the next. This set of real numbers is neither finite nor countably infinite. Any time that you ask yourself the question, "What are the possible values for the random variable?" and are forced to admit that the set of possibilities includes some interval or continuous span of real numbers, then the random variable being studied is not discrete.

3.2 DISCRETE PROBABILITY DENSITIES

When dealing with a random variable, it is not enough just to determine what values are possible. We also need to determine what is probable. We must be able to predict in some sense the values that the variable is likely to assume at any time. Since the behavior of a random variable is governed by chance, these predictions must be made in the face of a great deal of uncertainty. The best that can be done is to describe the behavior of the random variable in terms of probabilities. Two functions are used to accomplish this. We shall refer to these as the *density function* and the *cumulative distribution function*. The former is known by a variety of names in the discrete case, some of the most commonly encountered ones being the probability

function, the probability mass function, and the probability density function. In the discrete case, the density is denoted by either $p(x)$ or $f(x)$; in the continuous case it is almost always denoted by $f(x)$. For consistency we shall use $f(x)$ for the density in both cases. We begin by defining the density function for discrete random variables.

Definition 3.2.1 (Discrete density). Let X be a discrete random variable. The function f given by

$$f(x) = P[X = x]$$

for x real is called the density function for X.

There are several facts to note concerning the density in the discrete case. First, f is defined on the entire real line, and for any given real number x, $f(x)$ is the probability that the random variable X assumes the value x. For example, $f(2)$ is the probability that the random variable X assumes the numerical value of 2. Second, since $f(x)$ is a probability, $f(x) \geq 0$ regardless of the value of x. Third, if we sum f over all values of X that occur with nonzero probability, the sum must be 1. The following two conditions are necessary and sufficient conditions for a function f to be a discrete density. That is, if a function satisfies both of these conditions then it can be viewed as representing the density for some discrete random variable; if it fails to satisfy both then it cannot be the density for any discrete random variable:

Necessary and Sufficient Conditions for a Function to be a Discrete Density

1. $f(x) \geq 0$
2. $\sum_{\text{all } x} f(x) = 1$

The next example illustrates these ideas.

Example 3.2.1. Consider the random variable Y, the number of cells exposed to antigen-carrying lymphocytes in the presence of polyethylene glycol to obtain the first fusion (see Example 3.1.2). We know that under these conditions the probability that a given cell will fuse is 1/2. Thus the probability that it will not fuse is also 1/2. It is reasonable to assume that the cells behave independently. The possible values for Y are $\{1, 2, 3, \ldots\}$. The probability that the first cell will fuse is 1/2. That is,

$$P[Y = 1] = f(1) = 1/2$$

The probability that the first cell will not fuse but the second one will, yielding a value of 2 for Y, is

$$P[Y = 2] = f(2) = P[\text{first cell does not fuse}]P[\text{second cell does fuse}]$$
$$= 1/2 \cdot 1/2 = 1/4$$

Similarly,

$$P[Y = 3] = f(3) = 1/2 \cdot 1/2 \cdot 1/2 = 1/8$$

We can summarize the entire probability structure for Y in a density table (see Table 3.1). This is a table giving the possible values for the random variable in the first row and their corresponding probabilities in the second row. Note that there is an obvious pattern to the entries in row 2. When this occurs, we can find a closed-form expression for the density. In this case

$$f(y) = \begin{cases} (1/2)^y & y = 1, 2, 3, \ldots \\ 0 & \text{elsewhere} \end{cases}$$

Is this really a density? This function is obviously nonnegative, but does it sum to 1? To see this, note that

$$\sum_{\text{all } y} f(y) = \sum_{y=1}^{\infty} (1/2)^y$$

is a geometric series with first term $a = 1/2$ and common ratio $r = 1/2$. The properties of geometric series are well known. In particular, recall from elementary calculus that such a series can converge or diverge. The following fact will be useful in the material that follows:

Convergence of geometric series

Let $\sum_{k=1}^{\infty} ar^{k-1}$ be a geometric series.

The series converges to $\dfrac{a}{1-r}$ provided $|r| < 1$.

If we apply this result here, we see that

$$\sum_{y=1}^{\infty} (1/2)^y = \frac{a}{1-r} = \frac{1/2}{1-1/2} = 1$$

and the function f is a density.

Even though a discrete density is defined on the entire real line, it is only necessary to specify the density for those values y for which $f(y) \neq 0$. For instance, in the previous example we can write

$$f(y) = (1/2)^y \qquad y = 1, 2, 3, \ldots$$

It is understood that $f(y) = 0$ for all other real numbers.

Once it is known that a function is a density, it can be used to answer questions concerning the behavior of Y.

TABLE 3.1

y	1	2	3	4 · · ·
$P[Y = y] = f(y)$	1/2	1/2 · 1/2	1/2 · 1/2 · 1/2	1/2 · 1/2 · 1/2 · 1/2 · · ·

Example 3.2.2. What is the probability that we will need to expose four or more cells to antigen-carrying lymphocytes in the presence of polyethylene glycol to obtain the first fusion? That is, what is $P[Y \geq 4]$? The density for Y is

$$f(y) = (1/2)^y \qquad y = 1, 2, 3, \ldots$$

Although the desired probability can be found directly, it is easier to use subtraction:

$$\begin{aligned} P[Y \geq 4] &= 1 - P[Y < 4] \\ &= 1 - P[Y \leq 3] \\ &= 1 - (P[Y = 1] + P[Y = 2] + P[Y = 3]) \\ &= 1 - (f(1) + f(2) + f(3)) \\ &= 1 - ((1/2)^1 + (1/2)^2 + (1/2)^3) \\ &= 1 - (1/2 + 1/4 + 1/8) \\ &= 1 - 7/8 = 1/8 \end{aligned}$$

Cumulative Distribution

The second function used to compute probabilities is the cumulative distribution function F. Most of the statistical tables used in the material that follows are tables of the cumulative distribution function for some pertinent random variable.

The word "cumulative" suggests the role of this function. It sums or accumulates the probabilities found by means of the density. This function is defined as follows:

Definition 3.2.2 (Cumulative distribution—discrete). Let X be a discrete random variable with density f. The cumulative distribution function for X, denoted by F, is defined by

$$F(x) = P[X \leq x] \qquad \text{for } x \text{ real}$$

Consider a specific real number x_0. To find $P[X \leq x_0] = F(x_0)$, we sum the density f over all values of X that occur with nonzero probability that are less than or equal to x_0. That is, computationally,

$$F(x_0) = \sum_{x \leq x_0} f(x)$$

This idea is illustrated in Example 3.2.3.

Example 3.2.3. Certain genes produce such a tremendous deviation from normal that the organism is unable to survive. Such genes are called lethal genes. An example is the gene that produces a yellow coat in mice, Y. This gene is dominant over that for gray, y. Normal genetic theory predicts that when two yellow mice heterozygous for this trait (Yy) mate, 1/4 of the offspring will be gray and 3/4 will be yellow. Biologists have observed that these predicted proportions do not, in fact, occur, but that the actual

TABLE 3.2

x	0	1	2	3
$P[X = x] = f(x)$	1/27	6/27	12/27	8/27

TABLE 3.3

x	0	1	2	3
$P[X \le x] = F(x)$	1/27	7/27	19/27	27/27

TABLE 3.4

y	1	2	3	$4 \cdots$
$P[Y \le y] = F(y)$	8/16	12/16	14/16	$15/16 \cdots$

percentages produced are 1/3 gray and 2/3 yellow. It has been established that this shift is caused by the fact that 1/4 of the embryos, those homozygous for yellow (YY), do not develop. This leaves only two genotypes, Yy and yy, occurring in a ratio of 2 to 1, with the former producing a mouse with a yellow coat. For this reason, the gene Y is said to be lethal.

The density for X, the number of yellow mice in a litter of size 3, is shown in Table 3.2, and its cumulative distribution is given in Table 3.3. Notice that

$$F(0) = P[X \le 0] = P[X = 0] = 1/27$$
$$F(1) = P[X \le 1] = P[X = 0] + P[X = 1] = 1/27 + 6/27$$
$$F(2) = P[X \le 2] = P[X = 0] + P[X = 1] + P[X = 2]$$
$$= 1/27 + 6/27 + 12/27$$
$$F(3) = P[X \le 3] = 1$$

For discrete random variables that can assume only a finite number of possible values, the last entry in the bottom row of the cumulative table will always be 1.

Although cumulative probabilities are often given in table form as in the preceding example, it is sometimes possible to find express F in equation form. Example 3.2.4 illustrates this idea.

Example 3.2.4. Consider the random variable Y of Example 3.2.1 with density

$$f(y) = (1/2)^y \qquad y = 1, 2, 3, \ldots$$

A partial cumulative table for Y is shown in Table 3.4. It is formed by summing the probabilities given in the density table, Table 3.1. It is helpful to have a closed-form expression for F. In this case it is easy to obtain such an expression. By definition,

$$F(y_0) = \sum_{y \le y_0} f(y)$$

If we let $[y_0]$ denote the greatest integer less than or equal to y_0, then in this case $F(y_0)$ can be expressed as

$$F(y_0) = \sum_{y=1}^{[y_0]} (1/2)^y$$
$$= \sum_{y=1}^{[y_0]} (1/2)(1/2)^{y-1}$$

Recall from elementary calculus that the sum of the first n terms of a geometric series is given by

Sum of first n terms: Geometric series

$$\sum_{k=1}^{n} ar^{k-1} = \frac{a(1 - r^n)}{1 - r} \qquad r \ne 1$$

where a is the first term of the series and r is the common ratio.

Apply this result with $a = 1/2$ and $r = 1/2$, to obtain

$$F(y_0) = \frac{(1/2)[1 - (1/2)^{[y_0]}]}{1 - 1/2}$$
$$= 1 - (1/2)^{[y_0]}$$

The probability that at most seven cells must be exposed to obtain the first fusion is given by

$$P[Y \le 7] = F(y) = 1 - (1/2)^7 = \frac{127}{128}$$

3.3 EXPECTATION AND DISTRIBUTION PARAMETERS

The density function of a random variable completely describes the behavior of the variable. However, associated with any random variable are constants, or "parameters," that are descriptive. Knowledge of the numerical values of these parameters gives the researcher quick insight into the nature of the variables. We consider three such parameters: the mean μ, the variance σ^2, and the standard deviation σ. If the exact density of the random variable is known, then the numerical value of each parameter can be found from mathematical considerations. That is the topic of this section. If the only thing available to the researcher is a set of observations on the random variable (a data set), then the values of these parameters cannot be found exactly. They must be approximated by using statistical techniques. That is the topic of much of the remainder of this text.

To understand the reasoning behind most statistical methods, it is necessary to become familiar with one general concept, namely, the idea of *mathematical expectation* or *expected value*. This concept is used in defining many statistical parameters

and provides the logical basis for most of the methods of statistical inference presented later in this text.

A simple example will illustrate the basic idea of expectation. Consider the roll of a single fair die, and let X denote the number that is obtained. The possible values for X are 1, 2, 3, 4, 5, 6, and since the die is fair, the probability associated with each value is 1/6. The density for X is given by

$$f(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

When we ask for the expected value of X, we are asking for the *long-run theoretical average value of X*. If we imagine rolling the die over and over and recording the value of X for each roll, then we are asking for the theoretical average value of the rolls as the number of rolls approaches infinity. Since the density for X is symmetric and known, this average can be found intuitively. Notice that since $P[X = 1] = P[X = 6] = 1/6$, in the long run we expect to roll as many 1's as 6's. These values should counterbalance one another, and their average value is $(6 + 1)/2 = 3.5$. We also expect to roll as many 2's as 4's; these numbers also average to 3.5. Likewise, the numbers 3 and 4 are expected to counterbalance one another; they average 3.5. Logic dictates that, in the long run the average or expected value of X is 3.5. We write this as $E[X] = 3.5$. Notice that this value can be calculated from the density for X as follows:

$$E[X] = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6 = 3.5$$

or

$$E[X] = \sum_{\text{all } x} (\text{value of } x)(\text{probability})$$

Of course, the characteristic that makes finding this expectation easy is the symmetry of the density. Can we develop a definition of expectation that will work for nonsymmetric densities and that will apply not only to X, but also to random variables that are functions of X? The answer is "yes," and the desired definition is given in Definition 3.3.1. Let us point out that in most problems interest centers first on $E[X]$. However, expectations for functions of X such as X^2, $(X - c)^2$, where c is a constant and e^{tX} are especially useful in statistical theory. For this reason, the definition of expected value is given in general terms. We now define what we mean by the expected value of some function of X which we denote by $H(X)$.

Definition 3.3.1 (Expected value). Let X be a discrete random variable with density f. Let $H(X)$ be a random variable. The expected value of $H(X)$, denoted by $E[H(X)]$, is given by

$$E[H(X)] = \sum_{\text{all } x} H(x)f(x)$$

provided $\Sigma_{\text{all } x}|H(x)|\, f(x)$ is finite. Summation is over all values of X that occur with nonzero probability.

Note that in the special case in which $H(X) = X$, we obtain the expected value of X from this definition. Thus we see that

Expected Value of X

$$E[X] = \sum_{\text{all } x} x f(x)$$

One other thing to note concerning this definition is the fact the restriction that $\Sigma_{\text{all } x}|H(x)|\, f(x)$ exists is not particularly restrictive in practice. If the set of possible values for X is finite, it *will* be satisfied; if the set of possible values for X is countably infinite, it will *usually* be satisfied. However, it is possible to concoct a density f and a function $H(X)$ for which the series $\Sigma_{\text{all } x}|H(x)|\, f(x)$ does not converge. (See Exercise 22.) In this case we say that the expected value of the random variable $H(X)$ does not exist. An example will illustrate the use of Definition 3.3.1. Please realize that the density has been greatly oversimplified for purposes of illustration!

Example 3.3.1. A drug is used to maintain a steady heart rate in patients who have suffered a mild heart attack. Let X denote the number of heartbeats per minute obtained per patient. Consider the hypothetical density given in Table 3.5. What is the average heart rate obtained by all patients receiving this drug? That is, what is $E[X]$? By Definition 3.3.1,

$$\begin{aligned} E[X] &= \sum_{\text{all } x} H(x)f(x) \\ &= \sum_{\text{all } x} xf(x) \\ &= 40(.01) + 60(.04) + 68(.05) + \cdots + 100(.01) \\ &= 70 \end{aligned}$$

Since the number of possible values for X is finite, $\Sigma_{\text{all } x}|x|\, f(x)$ exists. Thus we can say that the *average* heart rate obtained by patients using this drug is 70 heartbeats per minute. Intuitively, we should have expected this result. Notice the symmetry of the density. In the long run we would expect as many patients with heart rates of 100 as with heart rates of 40; as many with a rate of 60 as with a rate of 80. Similarly, the rates of 68 and 72 occur with the same frequency. Each of these pairs averages to 70, the value obtained by the remaining 80% of the patients. Common sense points to 70 as the expected value for X.

When used in a statistical context, the expected value of a random variable X is referred to as its *mean* and is denoted by μ or μ_X. That is, the terms *expected*

TABLE 3.5

x	40	60	68	70	72	80	100
$f(x)$	.01	.04	.05	.80	.05	.04	.01

value and *mean* are interchangeable, as are the symbols $E[X]$ and μ. The mean can be thought of as a measure of the "center of location" in the sense that it indicates where the "center" of the density lies. For this reason, the mean is often referred to as a "location" parameter. To emphasize these points, let us summarize the preceding discussion.

Notes on the Expected Value of a Random Variable X

1. The expected value of a random variable is its theoretical average value. It is denoted by $E[X]$ and can be calculated from knowledge of the density for X.
2. In a statistical setting, the average value of X is called its *mean value.* Hence the terms *average value, mean value,* and *expected value* are interchangeable.
3. The mean value of X is denoted by the Greek symbol μ (*mu*). Hence the symbols μ and $E[X]$ are interchangeable.
4. The mean or expected value of X is one measure of the location of the center of the X values. For this reason, μ is called a "location" parameter.

There are three rules for handling expected values that are useful in justifying statistical procedures in later chapters. These rules hold for both continuous and discrete random variables. The rules are stated and illustrated here. We outline the proofs of the first two as exercises; the proof of rule 3 must be deferred until Chap. 5.

Theorem 3.3.1 (Rules for expectation). Let X and Y be random variables and let c be any real number.

1. $E[c] = c$ (The expected value of any constant is that constant.)
2. $E[cX] = cE[X]$ (Constants can be factored from expectations.)
3. $E[X + Y] = E[X] + E[Y]$ (The expected value of a sum is equal to the sum of the expected values.)

Example 3.3.2. Let X and Y be random variables with $E[X] = 7$ and $E[Y] = -5$. Then

$$\begin{aligned} E[4X - 2Y + 6] &= E[4X] + E[-2Y] + E[6] && \text{Rule 3} \\ &= 4E[X] + (-2)E[Y] + E[6] && \text{Rule 2} \\ &= 4E[X] - 2E[Y] + 6 && \text{Rule 1} \\ &= 4(7) - 2(-5) + 6 \\ &= 44 \end{aligned}$$

Variance and Standard Deviation

Knowledge of the mean of a random variable is important, but this knowledge *alone* can be misleading. The next example should show you the problem.

Example 3.3.3. Suppose that we wish to compare a new drug to that of Example 3.3.1. Let X denote the number of heartbeats per minute obtained using the old drug and Y the number per minute obtained with the new drug. The hypothetical density of

TABLE 3.6

x	40	60	68	70	72	80	100
$f(x)$	.01	.04	.05	.80	.05	.04	.01

y	40	60	68	70	72	80	100
$f(y)$	.40	.05	.04	.02	.04	.05	.40

each of these variables is given in Table 3.6. Since each of the densities is symmetric, inspection shows that $\mu_X = \mu_Y = 70$. Each drug produces *on the average* the same number of heartbeats per minute. However, there is obviously a drastic difference between the two drugs that is not being detected by the mean. The old drug produces fairly consistent reactions in patients, with 90% differing from the mean by at most 2; very few (2%) have an extreme reaction to the drug. However, the new drug produces highly diverse responses. Only 10% of the patients have heart rates within 2 units of the mean, whereas 80% show an extreme reaction. If we examined only the mean, we would conclude that the two drugs had identical effects—but nothing could be further from the truth!

It is obvious from Example 3.3.3 that something is not being measured by the mean. That something is *variability.* We must find a parameter that reflects consistency or the lack of it. We want the measure to assume a large positive value if the random variable fluctuates in the sense that it often assumes values far from its mean; the measure should assume a small positive value if the values of X tend to cluster closely about the mean. There are several ways to define such a measure. The most widely used is the *variance.*

Definition 3.3.2 (Variance). Let X be a random variable with mean μ. The variance of X, denoted by Var X, or σ^2, is given by

$$\text{Var } X = \sigma^2 = E[(X - \mu)^2]$$

Note that the variance measures variability by considering $X - \mu$, the difference between the variable and its mean. The difference is squared so that negative values will not cancel positive ones in the process of finding the expected value. When expressed in the form $E[(X - \mu)^2]$, it is easy to see that σ^2 has the properties that we want. When the variable X often assumes values far from μ, σ^2 will be a large positive number; when the values of X tend to fall close to μ, σ^2 will assume a small positive value. Figure 3.1 illustrates the idea.

Usually, the definition of σ^2 is not used to compute the variance. Rather, we use an alternative form which is given in the following theorem.

Theorem 3.3.2 (Computational formula for σ^2)

$$\sigma^2 = \text{Var } X = E[X^2] - (E[X])^2$$

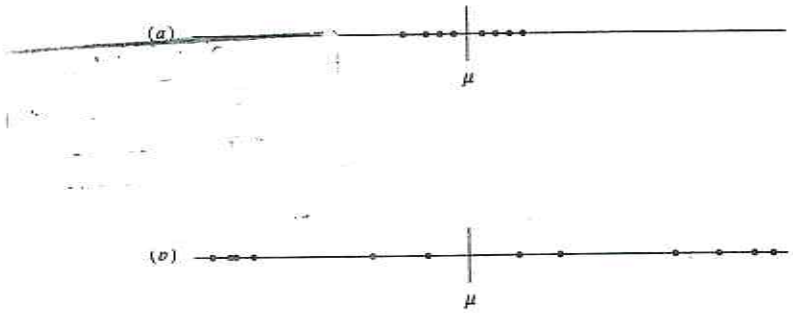


FIGURE 3.1
(*a*) A distribution with a small variance. Most of the data points, denoted by dots, lie fairly close to the average value, μ. Hence most of the differences, $x - \mu$, will be small; (*b*) a distribution with a large variance. Many of the data points lie far from the average value, μ.

Proof. By definition

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

Using the rules of expectation, Theorem 3.3.1, we obtain

$$\text{Var } X = E[X^2] - 2\mu E[X] + \mu^2$$

Since the symbols μ and $E[X]$ are interchangeable,

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

We illustrate the theorem by computing the variance of each of the random variables of Example 3.3.3.

Example 3.3.4. To find σ_X^2 and σ_Y^2 for the variables of Example 3.3.3, we first use Table 3.6 to find $E[X^2]$ and $E[Y^2]$. We know that $E[X] = E[Y] = 70$.

$$\begin{aligned}E[X^2] &= \sum_{\text{all } x} x^2 f(x) \\ &= (40^2)(.01) + (60^2)(.04) + \cdots + (100^2)(.01) \\ &= 4926.4\end{aligned}$$

$$\begin{aligned}E[Y^2] &= \sum_{\text{all } y} y^2 f(y) \\ &= (40^2)(.40) + (60^2)(.05) + \cdots + (100^2)(.40) \\ &= 5630.32\end{aligned}$$

By Theorem 3.3.2,

$$\begin{aligned}\text{Var } X &= E[X^2] - (E[X])^2 \\ &= 4926.4 - 70^2 = 26.4 \\ \text{Var } Y &= E[Y^2] - (E[Y])^2 \\ &= 5630.32 - 70^2 = 730.32\end{aligned}$$

As expected, Var $Y >$ Var X. Even though the drugs produce the same mean number of heartbeats per minute, they do not behave in the same way. The new drug is not as consistent in its effect as the old.

Note that the variance of a random variable reported alone is not very informative. Is a variance of 26.4 large or small? Only when this value is compared to the variance of a similar variable does it take on meaning. Hence variances are used often for comparative purposes to choose between two variables that otherwise appear to be identical. Also, note that the variance of a random variable is essentially a pure number whose associated units are often physically meaningless. When this occurs, the unit can be omitted. For example, the unit associated with the variance of Example 3.3.4 is a "squared heartbeat." This makes little sense, so in this case variance can be reported with no unit attached. To overcome this problem, a second measure of variability is employed. This measure is the nonnegative square root of the variance, and it is called the *standard deviation.* It has the advantage of having associated with it the same units as the original data.

Definition 3.3.3 (Standard deviation). Let X be a random variable with variance σ^2. The standard deviation of X, denoted by σ, is given by

$$\sigma = \sqrt{\text{Var } X} = \sqrt{\sigma^2}$$

Example 3.3.5. The standard deviations of variables X and Y of Example 3.3.4 are, respectively,

$$\sigma_X = \sqrt{\text{Var } X} = \sqrt{26.4} = 5.14 \text{ heartbeats per minute}$$
$$\sigma_Y = \sqrt{\text{Var } Y} = \sqrt{730.32} = 27.02 \text{ heartbeats per minute}$$

To emphasize these points we present a brief summary of the important aspects of the standard deviation of a random variable X.

Properties of standard deviation

1. The standard deviation of X is defined as the nonnegative square root of its variance.
2. The standard deviation is denoted by σ, and the variance of X is denoted by σ^2.
3. A large standard deviation implies that the random variable X is rather inconsistent and somewhat hard to predict; a small standard deviation is an indication of consistency and stability.

4. Standard deviation is always reported in physical measurement units that match the original data. Variance is often unitless.

Just as there are three rules for expectation that help in simplifying complex expressions, so are there three rules for variance. These rules parallel those for expectation. Rules 1 and 2 can be proved by using the rules for expectation (see Exercise 20). The proof of rule 3 must be deferred until the notion of "independent random variables" has been formalized.

Theorem 3.3.3 (Rules for variance). Let X and Y be random variables and c any real number. Then

1. $\text{Var } c = 0$
2. $\text{Var } cX = c^2 \text{ Var } X$
3. If X and Y are independent, then $\text{Var}(X + Y) = \text{Var } X + \text{Var } Y$

(Two variables are independent if knowledge of the value assumed by one gives no clue to the value assumed by the other.)

Example 3.3.6. Let X and Y be independent with $\sigma_X^2 = 9$ and $\sigma_Y^2 = 3$. Then

$$\begin{aligned} \text{Var}[4X - 2Y + 6] &= \text{Var}[4X] + \text{Var}[-2Y] + \text{Var } 6 && \text{Rule 3} \\ &= 16 \text{ Var } X + 4 \text{ Var } Y + \text{Var } 6 && \text{Rule 2} \\ &= 16 \text{ Var } X + 4 \text{ Var } Y + 0 && \text{Rule 1} \\ &= 16(9) + 4(3) = 156 \end{aligned}$$

In this section we discussed three *theoretical* parameters associated with a random variable X. We showed not only how to determine their numerical values from knowledge of the density, but also how to interpret them physically. Keep these things in mind, for they play a major role in the study of statistical methods for analyzing experimental data.

3.4 GEOMETRIC DISTRIBUTION AND THE MOMENT GENERATING FUNCTION

In this section we consider two important topics. We introduce the first *family* of discrete random variables to be discussed in this text. Random variables are members of a family in the sense that each member of the family is characterized by a density function of the same mathematical form, differing only with respect to the numerical value of some pertinent parameter or parameters. This first family, called *geometric*, is used extensively in the areas of games of chance and in statistical quality control. It is named geometric because, as you will see, its theoretical properties are derived by applying the mathematical properties of the geometric series that you encountered in elementary calculus. The second topic is a discussion of the *moment generating function*. This is a function, derived from the density, that

allows one to calculate ordinary moments of a distribution easily. This in turn makes it possible to calculate the mean and variance of a random variable without having to use the definitions of these terms to do so. In many cases, this approach is much simpler than a direct calculation from the definition. The function also provides a fingerprint or a unique identifier for each distribution. This idea will be illustrated later in this section.

Geometric Distribution

We begin by considering the family of *geometric* random variables. As you shall see, you have already encountered some random variables of this type even though the name "geometric random variable" was not mentioned at the time.

Geometric random variables arise in practice in experiments characterized by the following properties:

Geometric properties

1. The experiment consists of a series of trials. The outcome of each trial can be classed as being either a "success" (s) or a "failure" (f). A trial with this property is called a *Bernoulli* trial.
2. The trials are identical and independent in the sense that the outcome of one trial has no effect on the outcome of any other. The probability of success, p, remains the same from trial to trial.
3. The random variable X denotes the number of trials needed to obtain the first success.

The sample space for an experiment such as that just described is

$$S = \{s, fs, ffs, fffs, \ldots\}$$

Since the random variable X denotes the number of trials needed to obtain the first success, X assumes the values 1, 2, 3, 4, ... To find the density for X, we look for a pattern. Note that

$$\begin{aligned} P[X = 1] &= P[\text{success on first trial}] = p \\ P[X = 2] &= P[\text{fail on first trial and succeed on second trial}] \end{aligned}$$

Since the trials are independent, the latter probability can be found by multiplying. That is,

$$\begin{aligned} P[X = 2] &= P[\text{fail on first trial and succeed on second trial}] \\ &= P[\text{fail on first trial}]P[\text{succeed on second trial}] \\ &= (1 - p)(p) \end{aligned}$$

Similarly,

$$\begin{aligned} P[X = 3] &= P[\text{fail on first trial and fail on second trial and succeed on third trial}] \\ &= (1 - p)(1 - p)(p) = (1 - p)^2 p \end{aligned}$$

TABLE 3.7

x	1	2	3	4	5	⋯
$f(x)$	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	$(1-p)^4p$	⋯

You should be able to see that the density for X is given by Table 3.7, where the probabilities given in row 2 of the table exhibit a definite pattern. This pattern can be expressed in closed form as

$$f(x) = (1-p)^{x-1}p \qquad x = 1, 2, 3, \ldots$$

We now define a geometric random variable as being any random variable with a density of this form.

Definition 3.4.1 (Geometric distribution). A random variable X is said to have a geometric distribution with parameter p if its density f is given by

$$f(x) = (1-p)^{x-1}p \qquad 0 < p < 1$$
$$x = 1, 2, 3, \ldots$$

The function f given in this definition is a density. It is obviously nonnegative. Furthermore,

$$\sum_{x=1}^{\infty} (1-p)^{x-1}p$$

is a geometric series with first term $a = p$ and common ratio $r = (1-p)$. Thus the series sums to

$$\frac{a}{1-r} = \frac{p}{1-(1-p)} = 1$$

as desired. From this argument the reason for the name "geometric" distribution should be apparent.

In Exercise 26 you are asked to verify that the general expression for the cumulative distributions function for a geometric random variable is

$$F(x) = 1 - q^{[x]}$$

where q is the probability of failure and $[x]$ is the greatest integer less than or equal to x.

Example 3.4.1. Random digits are integers selected from among {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} one at a time in such a way that at each stage in the selection process the integer chosen is just as likely to be one digit as any other. In simulation experiments it is often necessary to generate a series of random digits. This can be done in a number of ways, the most common being by means of a computerized random number generator.

In generating such a series, let X denote the number of trials needed to obtain the first zero. This experiment consists of a series of independent, identical trials with "success" being the generation of a zero. The probability of success is $p = 1/10$. Since X denotes the number of trials needed to obtain the first success, X is a geometric random variable. Its density is found by substituting the value 1/10 for p in the expression for f given in Definition 3.4.1. That is,

$$f(x) = (1-p)^{x-1}p \qquad x = 1, 2, 3, \ldots$$

or

$$f(x) = (9/10)^{x-1}1/10 \qquad x = 1, 2, 3, \ldots$$

The cumulative distribution function for X is given by

$$F(x) = 1 - (.9)^{[x]}$$

Finding the mean of a geometric random variable from the definition is tricky! Consider the next example.

Example 3.4.2. Let us find the mean of the random variable X, the number of trials needed to obtain a zero when generating a series of random digits. By Definition 3.3.1,

$$E[X] = \sum_{x=1}^{\infty} xf(x)$$
$$= \sum_{x=1}^{\infty} x(9/10)^{x-1}1/10$$

That is,

$$E[X] = 1/10 + 18/100 + 243/1000 + 2916/10{,}000 + \cdots$$

This series is not geometric. Consider the series $(9/10)E[X]$.

$$(9/10)E[X] = 9/100 + 162/1000 + 2187/10{,}000 + 26{,}244/100{,}000 + \cdots$$

Subtracting the latter from the former, we obtain

$$(1/10)E[X] = 1/10 + 9/100 + 81/1000 + 729/10{,}000 + \cdots$$

This series is geometric with first term 1/10 and common ratio 9/10. Thus

$$(1/10)E[X] = \frac{1/10}{1-9/10} = 1$$

or

$$E[X] = \frac{1}{1/10} = 10$$

Moment Generating Function

As we have seen, the two expectations $E[X]$ and $E[X^2]$ are very useful, as they allow us to find the mean and variance of the random variable. These, and other expectations

of the form $E[X^k]$ for k a positive integer, are examples of what are called *ordinary moments*. This term is defined as follows:

Definition 3.4.2 (Ordinary moments). Let X be a random variable. The k^{th} ordinary moment for X is defined as $E[X^k]$.

Thus $E[X] = \mu$ is the first ordinary moment for X; $E[X^2]$ is its second ordinary moment. The preceding example shows that finding ordinary moments, even the first moment, from the definition of expectation is not always easy. Fortunately, it is often possible to obtain a function, called the *moment generating function*, which will enable us to find these moments with less effort.

Definition 3.4.3 (Moment generating function). Let X be a random variable with density f. The moment generating function for X (m.g.f.) is denoted by $m_X(t)$ and is given by

$$m_X(t) = E[e^{tX}]$$

provided this expectation is finite for all real numbers t in some open interval $(-h, h)$.

Since each geometric random variable has a density of the same general form, it is possible to find a general expression for the moment generating function for such a variable. This expression is given in Theorem 3.4.1.

Theorem 3.4.1 (Geometric moment generating function). Let X be a geometric random variable with parameter p. The moment generating function for X is given by

$$m_X(t) = \frac{pe^t}{1 - qe^t} \qquad t < -\ln q$$

where $q = 1 - p$.

Proof. The density for X is given by

$$f(x) = q^{x-1}p \qquad x = 1, 2, 3, \ldots$$

By definition

$$\begin{aligned} m_X &= E[e^{tX}] \\ &= \sum_{\text{all } x} e^{tx} f(x) \\ &= \sum_{x=1}^{\infty} e^{tx} q^{x-1} p \\ &= pq^{-1} \sum_{x=1}^{\infty} (qe^t)^x \end{aligned}$$

The series on the right is a geometric series with first term qe^t and common ratio qe^t. Thus

$$\begin{aligned} m_X(t) &= pq^{-1}\left(\frac{qe^t}{1 - qe^t}\right) \\ &= \frac{pe^t}{1 - qe^t} \end{aligned}$$

provided $|r| = |qe^t| < 1$. Since the exponential function is nonnegative and $0 < q < 1$, this restriction implies that $qe^t < 1$. The inequality is solved for t as follows:

$$\begin{aligned} qe^t &< 1 \\ e^t &< 1/q \\ \ln e^t &< \ln 1/q \\ t &< \ln 1 - \ln q \\ t &< -\ln q \end{aligned}$$

The next theorem shows how the moment generating function can be used to generate ordinary moments for a random variable X. Its proof is based on the Maclaurin series expansion for e^z. Recall that this series is as follows:

Maclaurin Series Expansion for e^z

$$e^z = 1 + z + z^2/2! + z^3/3! + z^4/4! + \cdots$$

Theorem 3.4.2. Let $m_X(t)$ be the moment generating function for a random variable X. Then

$$\left.\frac{d^k m_X(t)}{dt^k}\right|_{t=0} = E[X^k]$$

Proof. To prove this theorem, let $z = tX$. The Maclaurin series expansion for e^{tX} is

$$e^{tX} = 1 + tX + (tX)^2/2! + (tX)^3/3! + (tX)^4/4! + \cdots$$

By taking the expected value of each side of this equation, we obtain

$$\begin{aligned} m_X(t) = E[e^{tX}] &= E[1 + tX + t^2X^2/2! + t^3X^3/3! + t^4X^4/4! + \cdots] \\ &= 1 + tE[X] + t^2/2!E[X^2] + t^3/3!E[X^3] + t^4/4!E[X^4] + \cdots \end{aligned}$$

Differentiating this series term by term with respect to t, we see that

$$\frac{dm_X(t)}{dt} = E[X] + tE[X^2] + t^2/2!E[X^3] + t^3/3!E[X^4] + \cdots$$

When this derivative is evaluated at $t = 0$, every term except the first becomes 0. Hence

$$\left.\frac{dm_X(t)}{dt}\right|_{t=0} = E[X]$$

Taking the second derivative of $m_X(t)$, we obtain

$$\frac{d^2m_X(t)}{dt^2} = E[X^2] + tE[X^3] + t^2/2!E[X^4] + \cdots$$

Evaluating this derivative at $t = 0$ yields

$$\left.\frac{d^2m_X(t)}{dt^2}\right|_{t=0} = E[X^2]$$

This procedure can be continued to show that

$$\left.\frac{d^km_X(t)}{dt^k}\right|_{t=0} = E[X^k]$$

for any positive integer k as desired.

Let us use the moment generating function to find a general expression for the mean and variance of a geometric distribution with parameter p.

Theorem 3.4.3. Let X be a geometric random variable with parameter p. Then

$$E[X] = 1/p \qquad \text{and} \qquad \text{Var } X = q/p^2$$

Proof. For a geometric random variable with parameter p

$$m_X(t) = \frac{pe^t}{1 - qe^t}$$

$$\frac{dm_X(t)}{dt} = \frac{(1 - qe^t)pe^t + pe^tqe^t}{(1 - qe^t)^2}$$

$$= \frac{pe^t}{(1 - qe^t)^2}$$

Evaluating this derivative at $t = 0$, we obtain

$$E[X] = \left.\frac{dm_X(t)}{dt}\right|_{t=0} = \frac{p}{(1 - q)^2}$$

$$= p/p^2$$

$$= 1/p$$

Taking the second derivative of $m_X(t)$, we obtain

$$\frac{d^2m_X(t)}{dt^2} = \frac{(1 - qe^t)^2pe^t + 2pe^t(1 - qe^t)qe^t}{(1 - qe^t)^4}$$

$$= \frac{pe^t(1 - qe^t)[(1 - qe^t) + 2qe^t]}{(1 - qe^t)^4}$$

$$= \frac{pe^t(1 + qe^t)}{(1 - qe^t)^3}$$

Evaluating this derivative at $t = 0$, we see that

$$E[X^2] = \left.\frac{d^2m_X(t)}{dt^2}\right|_{t=0} = \frac{p(1 + q)}{(1 - q)^3} = \frac{(1 + q)}{p^2}$$

Now

$$\text{Var } X = E[X^2] - (E[X])^2$$

$$= \frac{1 + q}{p^2} - \frac{1}{p^2}$$

$$= \frac{q}{p^2}$$

We illustrate the use of these theorems by finding the moment generating function, mean, and variance for the random variable of Example 3.4.1.

Example 3.4.3. Consider the random variable X, the number of trials needed to obtain the first zero when generating a series of random digits. Since this random variable is geometric with parameter $p = 1/10$,

$$m_X(t) = \frac{pe^t}{1 - qe^t} = \frac{(1/10)e^t}{1 - (9/10)e^t}$$

$$\mu = E[X] = 1/p = 10$$

$$\sigma^2 = \text{Var } X = q/p^2 = \frac{9/10}{(1/10)^2} = 90$$

Note that this value for μ agrees with that obtained in Example 3.4.2.

The importance of the moment generating function for a random variable is not completely evident at this time. It does give us a way to find general expressions for the mean and variance as well as for the ordinary moments of an entire family of random variables. As we shall see later, the moment generating function, when it exists, serves as a fingerprint that completely identifies the random variable under study. That is, if a distribution has a moment generating function then it is unique. Thus, to identify a distribution from its moment generating function we need only look for and recognize a pattern and then the distribution is evident. For example, if an unknown random variable has moment generating function

$$m_X(t) \frac{.4e^t}{1 - .6e^t}$$

then we know that the random variable follows a geometric distribution with $p = .4$, because the moment generating function assumes the general form

$$\frac{pe^t}{1 - qe^t}$$

which is the geometric fingerprint.

3.5 BINOMIAL DISTRIBUTION

The next distribution to be studied is the *binomial* distribution. Once again, you have already seen some binomial random variables even though they were not

labeled as such at the time. The theoretical basis for working with this distribution is the binomial theorem presented in most beginning algebra courses. The statement of this theorem is as follows:

Binomial theorem

For any two real numbers a and b and any positive integer n,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$$

where $\binom{n}{k}$ is given by $\dfrac{n!}{k!(n-k)!}$.

To recognize a situation that involves a binomial random variable, you must be familiar with the assumptions that underlie this distribution, which are as follows:

Binomial properties

1. The experiment consists of a *fixed* number, n, of Bernoulli trials, trials that result in either a "success" (s) or a "failure" (f).
2. The trials are identical and independent, and therefore the probability of success, p, remains the same from trial to trial.
3. The random variable X denotes the number of successes obtained in the n trials.

Once we realize that the binomial model is appropriate from the physical description of the experiment, we shall want to describe the behavior of the binomial random variable involved. To do so, we need to consider the density for the random variable. To get an idea of the general form for the binomial density, let us consider the case in which $n = 3$. The sample space for such an experiment is

$$S = \{fff, sff, fsf, ffs, ssf, sfs, fss, sss\}$$

Since the trials are independent, the probability assigned to each sample point is found by multiplying. For example, the probabilities assigned to the sample points *fff* and *sff* are $(1-p)(1-p)(1-p) = (1-p)^3$ and $p(1-p)(1-p) = p(1-p)^2$, respectively. The random variable X assumes the value 0 only if the experiment results in the outcome *fff*. That is,

$$P[X=0] = (1-p)^3$$

However, X assumes the value 1 if the experiment results in any one of the outcomes *sff*, *fsf*, or *ffs*. Thus

$$P[X=1] = 3 \cdot p(1-p)^2$$

Similarly,

$$P[X=2] = 3 \cdot p^2(1-p)$$

and

$$P[X=3] = p^3$$

It is evident that for $x = 0, 1, 2, 3$

$$P[X=x] = c(x)p^x(1-p)^{3-x}$$

where $c(x)$ denotes the number of sample points that correspond to x successes. Such a sample point is expressed as a permutation of three letters, with x of these being s's and the rest, $3-x$, of these being f's. Using the formula for the number of permutations of indistinguishable objects studied in Chap. 1, we see that

$$c(x) = \frac{3!}{x!(3-x)!} = \binom{3}{x}$$

Thus the density for this binomial random variable is given by

$$f(x) = \binom{3}{x} p^x(1-p)^{3-x} \qquad x = 0, 1, 2, 3$$

To generalize this idea to n trials, we replace 3 by n to obtain the expression

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

This suggests the formal definition of the binomial distribution.

Definition 3.5.1 (Binomial distribution). A random variable X has a binomial distribution with parameters n and p if its density is given by

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$
$$0 < p < 1$$

where n is a positive integer.

To see that the function given in this definition is a density, note that it is nonnegative. Furthermore, by applying the binomial theorem with $k = x$, $a = p$, and $b = 1-p$ it can be seen that

$$\sum_{x=0}^{n} \binom{n}{x} p^x(1-p)^{n-x} = [p + (1-p)]^n = 1$$

as desired.

Example 3.5.1. Recent studies of German air traffic controllers have shown that it is difficult to maintain accuracy when working for long periods of time on data display screens. A surprising aspect of the study is that the ability to detect spots on a radar screen decreases as their appearance becomes too rare. The probability of correctly identifying a signal is approximately .9 when 100 signals arrive per 30-minute period. This probability drops to .5 when only 10 signals arrive at random over a 30-minute period. The hypothesis is that unstimulated minds tend to wander. Let X denote the number of signals correctly identified in a 30-minute time span in which 10 signals

arrive. This experiment consists of a series of $n = 10$ independent and identical Bernoulli trials with "success" being the correct identification of a signal. The probability of success is $p = 1/2$. Since X denotes the number of successes in a fixed number of trials, X is binomial. Its density is found by letting $n = 10$ and $p = 1/2$ in the expression for f given in Definition 3.5.1. That is,

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n$$

or

$$f(x) = \binom{10}{x} (1/2)^x (1/2)^{10-x} \qquad x = 0, 1, 2, \ldots, 10$$

The next theorem summarizes other theoretical properties of the binomial distribution. Its proof is left as an exercise (Exercise 43).

Theorem 3.5.1. Let X be a binomial random variable with parameters n and p.

1. The moment generating function for X is given by
$$m_X(t) = (q + pe^t)^n \qquad q = 1 - p$$
2. $E[X] = \mu = np$
3. $\text{Var } X = \sigma^2 = npq$

Example 3.5.2. The random variable X, the number of radar signals properly identified in a 30-minute period, is a binomial random variable with parameters $n = 10$ and $p = 1/2$. The moment generating function for this random variable is

$$m_X(t) = (1/2 + 1/2e^t)^{10}$$

Its mean is $\mu = np = 10(1/2) = 5$, and its variance is $\sigma^2 = npq = 10(1/2)(1/2) = 10/4$.

In statistical studies we shall usually be interested in computing the probability that the random variable assumes certain values. This probability can be computed from the density function, f, or from the cumulative distribution function, F. Since the binomial distribution comes into play in such a wide variety of physical applications, tables of the cumulative distribution function for selected values of n and p have been compiled. Table I of App. A is one such table. That is, Table I gives the values of

$$F(t) = \sum_{x=0}^{[t]} \binom{n}{x} p^x (1-p)^{n-x}$$

for selected values of n and p, where $[t]$ represents the greatest integer less than or equal to t. Its use is illustrated in the following example.

Example 3.5.3 Let X denote the number of radar signals properly identified in a 30-minute time period in which 10 signals are received. Assuming that X is binomial

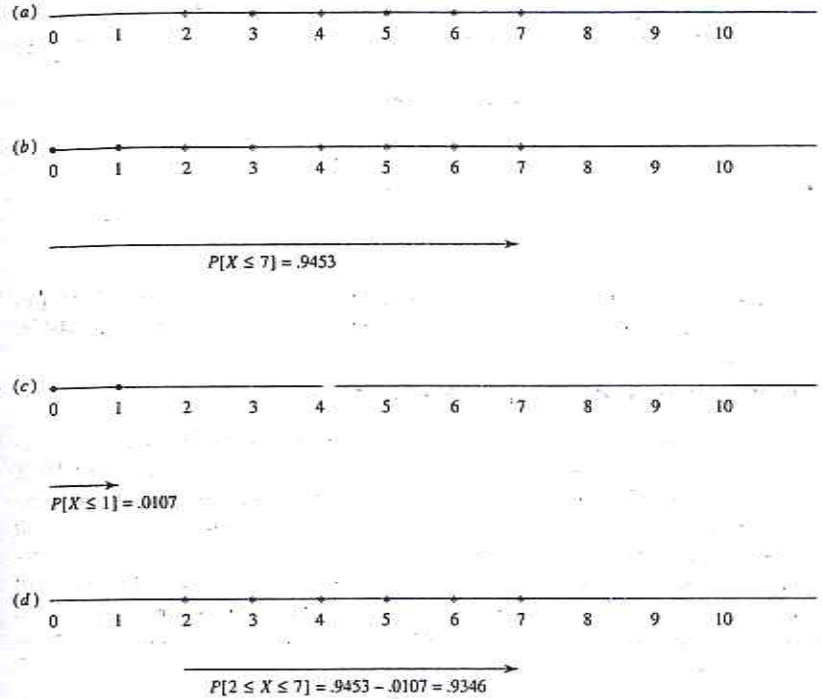


FIGURE 3.2
(*a*) The probability that X lies between 2 and 7 inclusive is the probability associated with the starred points; (*b*) $P[X \le 7] = .9453$ includes the probability associated with 0 and 1; (*c*) the probability associated with the unwanted points 0 and 1 is .0107; (*d*) the desired probability is found by subtraction.

with $n = 10$ and $p = 1/2$, find the probability that at most seven signals will be identified correctly. This probability can be found by summing the density from $x = 0$ to $x = 7$. That is,

$$P[X \le 7] = \sum_{x=0}^{7} \binom{10}{x} (1/2)^x (1/2)^{10-x}$$

Evaluating this probability directly entails a large amount of arithmetic. However, its value can be read from Table I of App. A. We first look at the group of values labeled $n = 10$. The desired probability of .9453 is found in the column labeled .5 and the row labeled 7. That is,

$$P[X \le 7] = F(7) = .9453$$

Other probabilities can be found. For example, find $P[2 \le X \le 7]$. Figure 3.2 suggests how this is done. Notice that in Fig. 3.2 we want the probability associated with points that are starred. To determine the desired probability, we first find the number 7 in Table I of App. A. Since the table is cumulative, the probability given, .9453, is the

We begin by considering the mathematical properties of this important family of random variables.

Definition 3.8.1 (Poisson distribution). A random variable X is said to have a Poisson distribution with parameter k if its density f is given by

$$f(x) = \frac{e^{-k}k^x}{x!} \qquad \begin{matrix} x = 0, 1, 2, \ldots \\ k > 0 \end{matrix}$$

The function f given in this definition is nonnegative. To see that it sums to 1, note that

$$\sum_{x=0}^{\infty} \frac{e^{-k}k^x}{x!} = e^{-k}(1 + k + k^2/2! + k^3/3! + \cdots)$$

The series on the right is the Maclaurin series for e^k. Thus

$$\sum_{x=0}^{\infty} \frac{e^{-k}k^x}{x!} = e^{-k}e^k = e^0 = 1$$

as desired.

The moment generating function for this distribution is easy to obtain, as is its mean and variance. The following theorem gives these results. Its proof is outlined as an exercise. (Exercise 69.)

Theorem 3.8.1. Let X be a Poisson random variable with parameter k.

1. The moment generating function for X is given by
$$m_X(t) = e^{k(e^t - 1)}$$
2. $E[X] = k$
3. $\text{Var } X = k$

Poisson random variables usually arise in connection with what are called *Poisson processes*. Poisson processes involve observing discrete events in a continuous "interval" of time, length, or space. We use the word "interval" in describing the general Poisson process with the understanding that we may not be dealing with an interval in the usual mathematical sense. For example, we might observe the number of white blood cells in a drop of blood. The discrete event of interest is the observation of a white cell, whereas the continuous "interval" involved is a drop of blood. We might observe the number of times radioactive gases are emitted from a nuclear power plant during a 3-month period. The discrete event of concern is the emission of radioactive gases. The continuous interval consists of a period of 3 months. The variable of interest in a Poisson process is X, the number of occurrences of the event in an interval of length s units. Although the derivation is a bit tricky, it can be shown using differential equations that X is a Poisson random

variable with parameter $k = \lambda s$, where λ is a positive number that characterizes the underlying Poisson process. To understand the physical significance of the constant λ, note that by Definition 3.8.1 the density for X is given by

$$f(x) = \frac{e^{-\lambda s}(\lambda s)^x}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

By Theorem 3.8.1 the expected value of X is λs. That is, the average number of occurrences of the event of interest in an interval of s units is λs. Thus the average number of occurrences of the event in 1 unit of time, length, area, or space is $\lambda s/s = \lambda$. That is, physically, the *parameter λ of a Poisson process represents the average number of occurrences of the event in question per measurement unit.*

The following steps are used in the solution of an applied Poisson problem:

Steps in Solving a Poisson Problem

1. Determine the basic unit of measurement being used.
2. Determine the average number of occurrences of the event per unit. This number is denoted by λ.
3. Determine the length or size of the observation period. This number is denoted by s.
4. The random variable X, the number of occurrences of the event in the interval of size s follows a Poisson distribution with parameter $k = \lambda s$.

These steps are illustrated in Example 3.8.1.

Example 3.8.1. The white blood cell count of a healthy individual can average as low as 6000 per cubic millimeter of blood. To detect a white-cell deficiency, a .001 cubic millimeter drop of blood is taken and the number of white cells X is found. How many white cells are expected in a healthy individual? If at most two are found, is there evidence of a white cell deficiency?

This experiment can be viewed as involving a Poisson process. The discrete event of interest is the occurrence of a white cell; the continuous interval is a drop of blood.

Let the measurement unit be a cubic millimeter; then $s = .001$ and λ, the average number of occurrences of the event per unit, is 6000. Thus X is a Poisson random variable with parameter $\lambda s = 6000(.001) = 6$. By Theorem 3.8.1, $E[X] = \lambda s = 6$. In a healthy individual we would expect, on the average, to see six white cells. How rare is it to see at most two? That is, what is $P[X \le 2]$? From Definition 3.8.1,

$$P[X \le 2] = \sum_{x=0}^{2} f(x) = \sum_{x=0}^{2} \frac{e^{-6}6^x}{x!}$$
$$= \frac{e^{-6}6^0}{0!} + \frac{e^{-6}6^1}{1!} + \frac{e^{-6}6^2}{2!}$$

Evaluating this type of expression directly does entail some arithmetic.

Once again, because of the wide appeal of the Poisson model, the values of the cumulative distribution function for selected values of the parameter $k = \lambda s$ are tabulated. Table II of App. A is one such table. The desired probability of .062 is found by

TABLE 3.8
Discrete distributions: A summary

Name	Density		Moment generating function	Mean	Variance
Geometric	$(1-p)^{x-1}p$	$x = 1, 2, 3, \ldots$ $0 < p < 1$	$\frac{pe^t}{1-qe^t}$	$\frac{1}{p}$	$\frac{q}{p^2}$
Uniform	$\frac{1}{n}$	$x = x_1, x_2, \ldots, x_n$ n a positive integer	$\frac{\sum_{i=1}^{n} e^{tx_i}}{n}$	$\frac{\sum_{i=1}^{n} x_i}{n}$	$\frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$x = 0, 1, 2, \ldots, n$ $0 < p < 1$ n a positive integer	$(q + pe^t)^n$	np	$np(1-p)$
Bernoulli or point binomial	$p^x(1-p)^{1-x}$	$x = 0, 1$ $0 < p < 1$	$q + pe^t$	p	$p(1-p)$
Hypergeometric	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$	$\max[0, n-(N-r)]$ $\leq x \leq \min(n, r)$		$n\frac{r}{N}$	$n\frac{r}{N}\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right)$
Negative binomial	$\binom{x-1}{r-1}(1-p)^{x-r}p^r$	$x = r, r+1, r+2, \ldots,$ $0 < p < 1$	$\frac{(pe^t)^r}{(1-qe^t)^r}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Poisson	$\frac{e^{-k}k^x}{x!}$	$x = 0, 1, 2, \ldots$ $k > 0$	$e^{k(e^t-1)}$	k	k

looking under the column labeled $k = 6$ in the row labeled 2. Is there evidence of a white-cell deficiency? There are no rules that say at what point probabilities are considered to be small. To answer this question, a value judgment must be made. If you consider .062 to be small, then the natural conclusion is that the individual does have a white-cell deficiency.

3.9 SIMULATING A DISCRETE DISTRIBUTION

In designing operating systems of various types, one often needs to simulate the system before it is built. Simulation is usually done with the aid of a computer. However, the idea behind simulation can be illustrated by using a random digit table. A portion of such a table is given in Table III of App. A. Its use is illustrated in the following example.

Example 3.9.1. Table 3.9 presents a portion of the random digit table in the appendix. Let us read a sequence of random two-digit numbers from this table. To do so, we must get a random start. This can be done by writing the integers 1 through 14 on slips of paper, placing the slips in a bowl, stirring, and drawing one slip at random from the bowl. The number selected identifies the column in which our starting number is located. In a similar way, we can select the row in which the starting number is located.

TABLE 3.9

Row	Column	Random digits (1)	(2)	(3)
1		10480	15011	01536
2		22368	46573	25595
3		24130	48360	22527
4		42167	93093	06243
5		37570	39975	81837
6		77921	06907	11008
7		99562	72905	56420
8		96301	91977	05463
9		89579	14342	63661
10		85485	36857	43342

Suppose that this process results in the selection of column 2 and row 5. This identifies the random starting point as 39975.

Since we want two-digit numbers, we need only read the first two digits of this number. Thus our first random number is 39. Since a random digit table is constructed in such a way that the digit appearing at each position in the table is just as likely to be one digit as any other, the table can be read in any way. Let us agree to read down the second column so that the next four two-digit numbers are 06, 72, 91, and 14.

The next example illustrates the use of a random digit table in a simple simulation experiment.

Example 3.9.2. Suppose that at a particular airport planes arrive at an average rate of one per minute and depart at the same average rate. We are interested in simulating the behavior of the random variable Z, the number of planes on the ground at a given time. We will simulate Z for five consecutive one-minute periods. Note that for each of these periods the random variables X, the number of arrivals, and Y, the number of departures, are both Poisson variables with parameter $k = 1$. The density for X and Y is obtained from Table II of App. A and is shown below:

$$P[X = 0] = P[Y = 0] = .368$$
$$P[X = 1] = P[Y = 1] = .368$$
$$P[X = 2] = P[Y = 2] = .184$$
$$P[X = 3] = P[Y = 3] = .061$$
$$P[X = 4] = P[Y = 4] = .015$$
$$P[X = 5] = P[Y = 5] = .003$$
$$P[X = 6] = P[Y = 6] = .001$$
$$P[X > 6] = P[Y > 6] \doteq 0$$

There are 1000 possible three-digit numbers. We divide them into seven categories to reflect the above probabilities. This division is shown in Table 3.10. To perform the simulation, we read a total of 10 random three-digit numbers using the procedure demonstrated in Example 3.9.1. Assume that at the beginning of the simulation there

CHAPTER 4

CONTINUOUS DISTRIBUTIONS

In Chap. 3 we learned to distinguish a discrete random variable from one that is not discrete. In this chapter we consider a large class of nondiscrete random variables. In particular, we consider random variables that are called *continuous*. We first study the general properties of variables of the continuous type and then present some important families of continuous random variables.

4.1 CONTINUOUS DENSITIES

In Chap. 3 we considered the random variable T, the time of the peak demand for electricity at a particular power plant. We agreed that this random variable is not discrete since, "a priori"—before the fact—we cannot limit the set of possible values for T to some finite or countably infinite collection of times. Time is measured continuously, and T can conceivably assume any value in the time interval [0, 24), where 0 denotes 12 midnight one day and 24 denotes 12 midnight the next day. Furthermore, if we ask *before* the day begins, What is the probability that the peak demand will occur exactly 12.013 278 650 931 271? the answer is 0. It is virtually impossible for the peak load to occur at this split second in time, not the slightest bit earlier or later. These two properties, possible values occurring as intervals and the a priori probability of assuming any specific value being 0, are the characteristics that identify a random variable as being continuous. This leads us to our next definition.

Definition 4.1.1 (Continuous random variable). A random variable is continuous if it can assume any value in some interval or intervals of real numbers and the probability that it assumes any specific value is 0.

Note that the statement that the probability that a continuous random variable assumes any specific value is 0 is essential to the definition. Discrete variables have no such restriction. For this reason, we calculate probabilities in the continuous case differently than we do in the discrete case. In the discrete case we defined a function f, called the density, which enabled us to compute probabilities associated with the random variable X. This function is given by

$$f(x) = P[X = x] \qquad x \text{ real}$$

This definition cannot be used in the continuous case because $P[X = x]$ is always 0. However, we do need a function that will enable us to compute probabilities associated with a continuous random variable. Such a function is also called a density.

Definition 4.1.2 (Continuous density). Let X be a continuous random variable. A function f such that

1. $f(x) \geq 0$ for x real
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P[a \leq X \leq b] = \int_a^b f(x)dx$ for a and b real

is called a density for X.

Although this definition may look arbitrary at first glance, it is not. Note that, as in the discrete case, f is defined over the entire real line and is nonnegative. Recall from elementary calculus that integration is the natural extension of summation in the sense that the integral is the limit of a sequence of Riemann sums. In the discrete case we require that $\sum_{\text{all } x} f(x) = 1$. The natural extension of this requirement to the continuous case is that the density integrate to 1. Therefore the necessary and sufficient conditions for a function to be a density for a continuous random variable are as follows:

Necessary and Sufficient Conditions for a Function to be a Continuous Density

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

In the discrete case we find the probability that X assumes a value in some set A by summing $f(x)$ over all values of x in A. That is,

$$P[X \in A] = \sum_{x \in A} f(x).$$

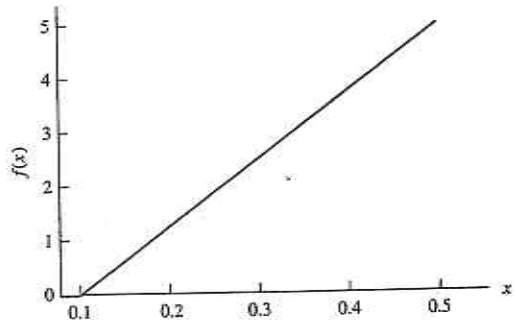


FIGURE 4.1
Graph of

$$f(x) = \begin{cases} 12.5x - 1.25 & .1 \le x \le .5 \\ 0 & \text{elsewhere} \end{cases}$$

In the continuous case we shall be interested in finding the probability that X assumes values in some interval $[a, b]$. Replacing A by $[a, b]$ and substituting integration for summation in the previous expression suggest property 3 of Definition 4.1.2. That is,

$$P[a \le X \le b] = \int_a^b f(x)dx$$

It is evident that the term "density" in the continuous case is just an extension of the ideas presented in the discrete case, with summation being replaced by integration. This is an important notion, as it will allow us to define the concept of expected value in the continuous case quite naturally.

Example 4.1.1. The lead concentration in gasoline currently ranges from .1 to .5 grams per liter. What is the probability that the lead concentration in a randomly selected liter of gasoline will lie between .2 and .3 grams inclusive? To answer this question, we need a density, f, for the random variable X, the number of grams of lead per liter of gasoline. Consider the function

$$f(x) = \begin{cases} 12.5x - 1.25 & .1 \le x \le .5 \\ 0 & \text{elsewhere} \end{cases}$$

The graph of f is shown in Fig. 4.1. The function is nonnegative. Furthermore,

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{.1}^{.5} (12.5x - 1.25)dx \\ &= \left[\frac{12.5x^2}{2} - 1.25x\right]_{.1}^{.5} \\ &= \left[\frac{12.5(.5)^2}{2} - 1.25(.5)\right] - \left[\frac{12.5(.1)^2}{2} - 1.25(.1)\right] \\ &= .9375 - (-.0625) = 1 \end{aligned}$$

Thus f satisfies properties 1 and 2 of Definition 4.1.2. Property 3 allows us to use f to find the desired probability. In particular,

$$\begin{aligned} P[.2 \le X \le .3] &= \int_{.2}^{.3} f(x)dx \\ &= \int_{.2}^{.3} (12.5x - 1.25)dx \\ &= \left[\frac{12.5x^2}{2} - 1.25x\right]_{.2}^{.3} \\ &= \left[\frac{12.5(.3)^2}{2} - 1.25(.3)\right] - \left[\frac{12.5(.2)^2}{2} - 1.25(.2)\right] \\ &= .1875 \end{aligned}$$

There are several important points to be made concerning the density in the continuous case. First, we shall follow the convention of defining f only over intervals for which $f(x)$ may be nonzero. For values of x not explicitly mentioned, $f(x)$ is assumed to be 0. In Example 4.1.1 we could have written f as

$$f(x) = 12.5x - 1.25 \qquad .1 \le x \le .5$$

with the understanding that $f(x) = 0$ elsewhere. Second, since the integral of a nonnegative function can be thought of as an area, properties 2 and 3 of Definition 4.1.2 can be expressed in terms of areas. In particular, property 2 requires that *the total area under the graph of f be 1.* Property 3 implies that the probability that the variable assumes a value between two points a and b is *the area under the graph of f between $x = a$ and $x = b$.* These ideas as they apply to Example 4.1.1 are demonstrated in Figs. 4.2(a) and (b), respectively. Third, since $P[X = a] = P[X = b] = 0$ in the continuous case,

$$P[a \le X \le b] = P[a \le X < b] = P[a < X \le b] = P[a < X < b].$$

In Example 4.1.1 the probability that the lead concentration in a liter of gasoline lies between .2 and .3 gram inclusive, $P[.2 \le X \le .3]$, is the same as $P[.2 < X < .3]$, the probability that it lies strictly between .2 and .3 gram. See Fig. 4.2(c). Fourth, properties 1 and 2 of Definition 4.1.2 are necessary and sufficient conditions for a function to be a density for a continuous random variable X. However, the density chosen for X cannot be just any function satisfying these conditions. It should be a function that assigns reasonable probabilities to events via property 3 of Definition 4.1.2. Whether or not the function f given in Example 4.1.1 satisfies this criteria is debatable. It was chosen for illustrative purposes only. Finding an appropriate density is not always easy. Some methods for helping in the selection of a density are discussed in Chap. 6.

Cumulative Distribution

The idea of a cumulative distribution function in the continuous case is useful. It is defined exactly as in the discrete case although found by using integration rather than summation.

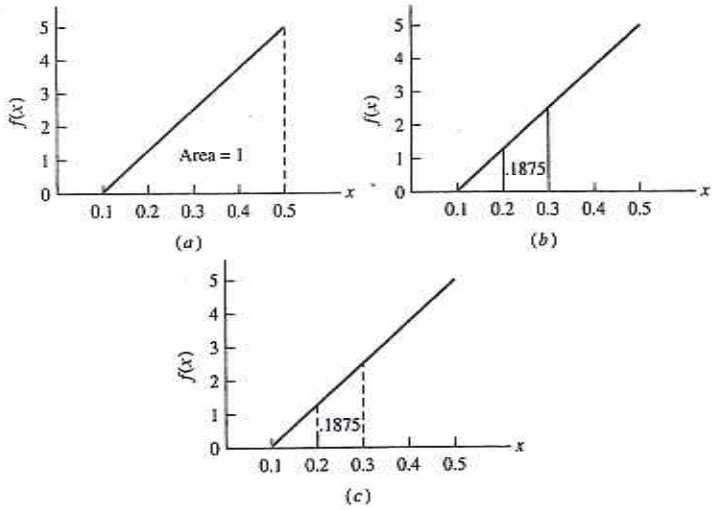


FIGURE 4.2
(*a*) $\int_{-\infty}^{\infty} f(x)dx = 1$ implies that the total area under the graph of f is 1; (*b*) $P[.2 \le X \le .3] = \int_{.2}^{.3}(12.5x - 1.25)dx = .1875$ implies that the area under the graph of f between $x = .2$ and $x = .3$ is .1875; (*c*) $P[.2 < X < .3] = P[.2 \le X \le .3] = .1875$.

Definition 4.1.3 (Cumulative distribution—continuous). Let X be continuous with density f. The cumulative distribution function for X, denoted by F, is defined by

$$F(x) = P[X \le x] \qquad x \text{ real}$$

To find $F(x)$ for a specific real number x, we integrate the density over all real numbers that are less than or equal to x.

Computing F Continuous Case

$$P[X \le x] = F(x) = \int_{-\infty}^{x} f(t)dt \qquad x \text{ real}$$

Graphically, this probability corresponds to the area under the graph of the density to the left of and including the point x.

Example 4.1.2. The density for the random variable X, the lead content in a liter of gasoline, is

$$f(x) = 12.5x - 1.25 \qquad .1 \le x \le .5$$

The cumulative distribution function for X is

$$P[X \le x] = F(x) = \int_{-\infty}^{x} f(t)dt$$

For $x < .1$ this integral has value 0 since for these values of x, $f(t)$ is itself 0. For $.1 \le x \le .5$,

$$\begin{aligned} F(x) = \int_{-\infty}^{x} f(t)dt &= \int_{.1}^{x}(12.5t - 1.25)dt \\ &= \left[\frac{12.5t^2}{2} - 1.25t\right]_{.1}^{x} \\ &= 6.25x^2 - 1.25x + .0625 \end{aligned}$$

For $x > .5$ the integral has value 1 since for these values of x we have integrated the density over its entire set of possible values. Summarizing, F is given by

$$F(x) = \begin{cases} 0 & x < .1 \\ 6.25x^2 - 1.25x + .0625 & .1 \le x \le .5 \\ 1 & x > .5 \end{cases}$$

What is the probability that the lead concentration in a randomly selected liter of gasoline will lie between .2 and .3 gram per liter? To answer this question, we rewrite it in terms of the cumulative distribution

$$\begin{aligned} P[.2 \le X \le .3] &= P[X \le .3] - P[X < .2] \\ &= P[X \le .3] - P[X \le .2] \qquad (X \text{ is continuous}) \\ &= F(.3) - F(.2) \end{aligned}$$

By substitution,

$$\begin{aligned} F(.3) &= 6.25(.3)^2 - 1.25(.3) + .0625 = .2500 \\ F(.2) &= 6.25(.2)^2 - 1.25(.2) + .0625 = .0625 \end{aligned}$$

Thus

$$\begin{aligned} P[.2 \le X \le .3] &= F(.3) - F(.2) \\ &= .2500 - .0625 = .1875 \end{aligned}$$

Note that this agrees with the result obtained in Example 4.1.1 using direct integration. Note also that $F(.3)$ gives the area to the left of .3 shown in Fig. 4.3(*a*); $F(.2)$ gives the area to the left of .2 shown in Fig. 4.3(*b*). When we form the difference $F(.3) - F(.2)$, we naturally obtain the area between .2 and .3 given in Fig. 4.3(*c*).

Recall that in the discrete case, the cumulative distribution, F, was obtained from the density by addition; if F was available, f could be obtained by subtraction, the operation that reverses addition. The same sort of thing happens in the continuous case. We obtain the cumulative distribution from the density by integrating f; if F is available, we can retrieve f by reversing the integration operation via differentiation. That is, in the continuous case,

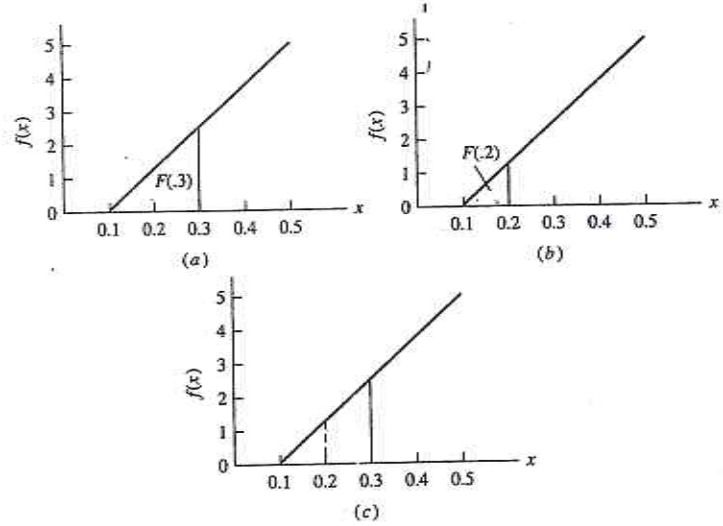


FIGURE 4.3
(a) $F(.3) = P[X \le .3]$; (b) $F(.2) = P[X \le .2]$; (c) $F(.3) - F(.2) = P[.2 \le X \le .3]$.

Obtaining f from F in the Continuous Case

$$f(x) = F'(x)$$

Example 4.1.3. In Example 4.1.2, we derived the cumulative distribution

$$F(x) = 6.25x^2 - 1.25x + .0625 \qquad .1 \le x \le .5$$

Note that

$$F'(x) = 12.5x - 1.25 \qquad .1 \le x \le .5$$

This is, as expected, the expression for the density for X that was given in Example 4.1.2.

Uniform Distribution

Perhaps the simplest continuous distribution with which to work is the *uniform* distribution. This distribution parallels the discrete uniform distribution presented in Exercise 34 of Chap. 3 in that, in a sense, events occur with equal or uniform probability. Since it is easy and instructive to develop the properties of this family of random variables directly from the definition, we leave the derivations to you. Important properties and applications are given in Exercises 5, 6, 10, 11, 18, and 19.

4.2 EXPECTATION AND DISTRIBUTION PARAMETERS

In this section we define the term *expected value for continuous random variables*. We also discuss how to use the definition to find the moment generating function, the mean, and the variance of a variable of the continuous type. As you will see, the definition parallels that given in the discrete case, with the summation operation being replaced by integration.

Definition 4.2.1 (Expected value). Let X be a continuous random variable with density f. Let $H(X)$ be a random variable. The expected value of $H(X)$, denoted by $E[H(X)]$, is given by

$$E[H(X)] = \int_{-\infty}^{\infty} H(x)f(x)dx$$

provided

$$\int_{-\infty}^{\infty} |H(x)|\, f(x)dx$$

is finite.

As in the discrete case, the mean or expected value of X is a special case of the above definition.

Expected Value of X

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

We illustrate the use of this definition by finding the mean and variance of the random variable X of Example 4.1.1. Recall that, by Theorem 3.3.2, the variance for X can be found via the computational shortcut

$$\sigma^2 = \text{Var}\,(X) = E[X^2] - (E[X])^2$$

Example 4.2.1. The density for X, the lead concentration in gasoline in grams per liter, is given by

$$f(x) = 12.5x - 1.25 \qquad .1 \le x \le .5$$

The mean or expected value of X is

$$\begin{aligned}
\mu = E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\
&= \int_{.1}^{.5} x(12.5x - 1.25)dx \\
&= \left[\frac{12.5x^3}{3} - \frac{1.25x^2}{2}\right]_{.1}^{.5} \\
&= \left[\frac{(12.5)(.5)^3}{3} - \frac{1.25(.5)^2}{2}\right] - \left[\frac{12.5(.1)^3}{3} - \frac{1.25(.1)^2}{2}\right] \\
&\doteq .3667 \text{ g/liter}
\end{aligned}$$

Since integration is over an interval of finite length

$$\int_{-\infty}^{\infty} |x| f(x)\,dx$$

exists. We can conclude that, on the average, a liter of gasoline contains approximately .3667 g of lead. How much variability is there from liter to liter? To answer this question, we find $E[X^2]$ and apply Theorem 3.3.2 to find the variance of X:

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\,dx \\ &= \int_{.1}^{.5} x^2(12.5x - 1.25)\,dx \\ &= \left[\frac{12.5x^4}{4} - \frac{1.25x^3}{3}\right]_{.1}^{.5} \doteq .1433 \end{aligned}$$

By Theorem 3.3.2,

$$\text{Var } X = E[X^2] - (E[X])^2 \doteq .1433 - (.3667)^2 \doteq .00883$$

The standard deviation of X is

$$\sigma = \sqrt{\text{Var } X} = \sqrt{.00883} = .09396 \text{ g/liter}$$

As in the discrete case, the moment generating function for a continuous random variable X is defined as $E[e^{tX}]$ provided this expectation exists for t in some open interval about 0. Its use is illustrated in the following example.

Example 4.2.2. The spontaneous flipping of a bit stored in a computer memory is called a "soft fail." Let X denote the time in millions of hours before the first soft fail is observed. Suppose that the density for X is given by

$$f(x) = e^{-x} \qquad x > 0$$

The mean and variance for X can be found directly using the method of Example 4.2.1. However, to find $E[X]$ and $E[X^2]$, integration by parts is required. This method of integration, although not difficult, is time-consuming. Let us find the moment generating function for X and use it to compute the mean and variance. By definition,

$$m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\,dx$$

In this case,

$$\begin{aligned} m_X(t) &= \int_0^{\infty} e^{tx} e^{-x}\,dx \\ &= \int_0^{\infty} e^{(t-1)x}\,dx \\ &= \frac{1}{t-1} e^{(t-1)x}\Big|_0^{\infty} \end{aligned}$$

Assume that $|t| < 1$. This guarantees that the exponent $(t-1)x < 0$, allowing us to evaluate the above integral. In particular,

$$m_X(t) = \frac{1}{1-t} \qquad |t| < 1$$

Since $e^{tx} > 0$, $|e^{tx}| = e^{tx}$. Thus the above argument has shown that

$$\int_{-\infty}^{\infty} |e^{tx}| f(x)\,dx$$

exists, as required in Definition 4.2.1. To use $m_X(t)$ to find $E[X]$ and $E[X^2]$, we apply Theorem 3.4.2. Note that

$$\begin{aligned} \frac{dm_X(t)}{dt} &= \frac{d(1-t)^{-1}}{dt} = (1-t)^{-2} \\ \frac{d^2 m_X(t)}{dt^2} &= 2(1-t)^{-3} \\ E[X] &= \frac{dm_X(t)}{dt}\Big|_{t=0} = 1 \\ E[X^2] &= \frac{d^2 m_X(t)}{dt^2}\Big|_{t=0} = 2 \\ \text{Var } X &= E[X^2] - (E[X])^2 = 2 - 1^2 = 1 \end{aligned}$$

The average or mean time that one must wait to observe the first soft fail is 1 million hours. The variance in waiting time is 1, and the standard deviation is 1 million hours.

To find the distribution parameters μ, σ^2, and σ, we can use either Definition 4.2.1 or the moment generating function technique. In practice, use whichever method is easier.

It should be pointed out that there is a nice geometric interpretation of the mean in the case of a continuous random variable. Imagine cutting out of a piece of thin rigid metal the region bounded by the graph of f and the x axis, and attempting to balance this region on a knife-edge held parallel to the vertical axis. The point at which the region would balance, if such a point exists, is the mean of X. Thus, μ_X is a "location" parameter in that it indicates the position of the center of the density along the x axis. The variance can also be interpreted pictorially. In the continuous case variance is a "shape" parameter in the sense that a random variable with small variance will have a compact density; one with a large variance will have a density that is rather spread out or flat.

4.3 GAMMA, EXPONENTIAL, AND CHI-SQUARED DISTRIBUTIONS

In this section we consider the gamma distribution. This distribution is especially important in that it allows us to define two families of random variables, the exponential and chi-squared, that are used extensively in applied statistics. The theoretical basis for the gamma distribution is the gamma function, a mathematical function defined in terms of an integral.

Definition 4.3.1 (Gamma function). The function Γ defined by

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1}e^{-z}dz \qquad \alpha > 0$$

is called the gamma function.

Theorem 4.3.1 presents two numerical properties of the gamma function that are useful in evaluating the function for various values of α. Its proof is outlined in Exercise 26.

Theorem 4.3.1 (Properties of the gamma function)

1. $\Gamma(1) = 1$.
2. For $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.

The use of Theorem 4.3.1 is illustrated in the next example.

Example 4.3.1

(*a*) Evaluate $\int_0^\infty z^3e^{-z}\,dz$. To evaluate this integral using the methods of elementary calculus requires repeated applications of integration by parts. To evaluate the integral quickly, rewrite it as

$$\int_0^\infty z^3e^{-z}dz = \int_0^\infty z^{4-1}e^{-z}dz$$

The integral on the right is $\Gamma(4)$. By applying Theorem 4.3.1 repeatedly, it can be seen that

$$\begin{aligned}\int_0^\infty z^3e^{-z}dz = \Gamma(4) &= 3\cdot\Gamma(3)\\ &= 3\cdot 2\cdot\Gamma(2)\\ &= 3\cdot 2\cdot 1\cdot\Gamma(1)\\ &= 3\cdot 2\cdot 1 = 6\end{aligned}$$

(*b*) Evaluate $\int_0^\infty (1/54)x^2e^{-x/3}\,dx$. To evaluate this integral, we make a change of variable, a technique that is used extensively in deriving the properties of the gamma distribution. In particular, let $z = x/3$ or $3z = x$. Then $3\,dz = dx$ and the problem becomes

$$\begin{aligned}\int_0^\infty (1/54)x^2e^{-x/3}dx &= \int_0^\infty 1/54(3z)^2e^{-z}3dz\\ &= 27/54\int_0^\infty z^2e^{-z}dz\end{aligned}$$

However,

$$\begin{aligned}\int_0^\infty z^2e^{-z}dz = \int_0^\infty z^{3-1}e^{-z}dz &= \Gamma(3)\\ &= 2\cdot\Gamma(2)\\ &= 2\cdot 1\cdot\Gamma(1)\\ &= 2\cdot 1 = 2\end{aligned}$$

Thus

$$\int_0^\infty (1/54)x^2e^{-x/3}dx = 27/54\cdot 2 = 1$$

Note that since the nonnegative function

$$f(x) = (1/54)x^2e^{-x/3}$$

has been shown to integrate to 1, it can be thought of as being a density for a continuous random variable X.

It is now possible to define the gamma distribution.

Gamma Distribution

Definition 4.3.2 (Gamma distribution). A random variable X with density

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta} \qquad \begin{array}{l}x > 0\\ \alpha > 0\\ \beta > 0\end{array}$$

is said to have a gamma distribution with parameters α and β.

Although the mean and variance of a gamma random variable can be found easily from the definitions of these parameters (see Exercise 31), we shall use the moment generating function technique. As you will see later, it is very helpful to know the form of the moment generating function for a random variable.

Theorem 4.3.2. Let X be a gamma random variable with parameters α and β. Then

1. The moment generating function for X is given by
$$m_X(t) = (1 - \beta t)^{-\alpha} \qquad t < 1/\beta$$
2. $E[X] = \alpha\beta$
3. $\operatorname{Var} X = \alpha\beta^2$

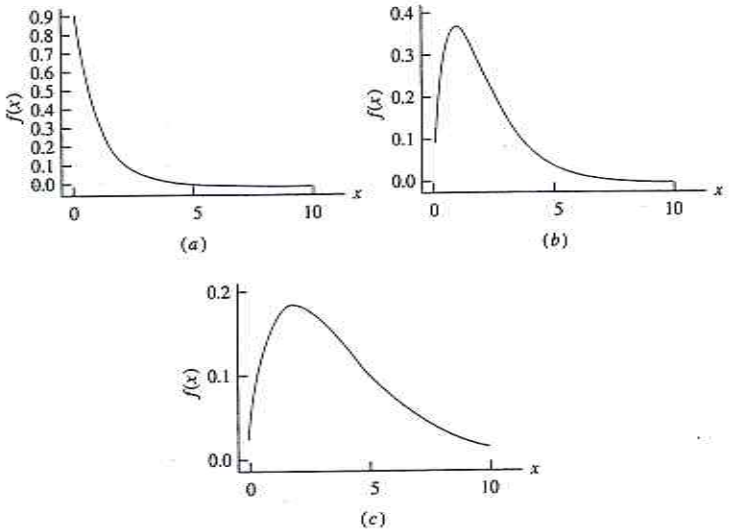


FIGURE 4.4
(a) $\alpha = 1, \beta = 1, \mu_X = 1, \sigma_X^2 = 1$; (b) $\alpha = 2, \beta = 1, \mu_X = 2, \sigma_X^2 = 2$; (c) $\alpha = 2, \beta = 2, \mu_X = 4, \sigma_X^2 = 8$.

The proof of this theorem is found in Appendix C.

Figure 4.4 shows the graphs of some gamma densities for a few values of α and β. Note that α and β both play a role in determining the mean and the variance of the random variable. Note also that the curves are not symmetric and are located entirely to the right of the vertical axis. It can be shown that for $\alpha > 1$, the maximum value of the density occurs at the point $x = (\alpha - 1)\beta$. (See Exercise 32.)

Exponential Distribution

As mentioned earlier, the gamma distribution gives rise to a family of random variables known as the *exponential* family. These variables are each gamma random variables with $\alpha = 1$. The density for an exponential random variable therefore assumes the form

Exponential density

$$f(x) = \frac{1}{\beta} e^{-x/\beta} \qquad x > 0$$
$$\beta > 0$$

The graph of a typical exponential density is shown in Fig. 4.4(*a*). This distribution arises often in practice in conjunction with the study of Poisson processes, which were discussed in Sec. 3.8. Recall that in a Poisson process discrete events are being observed over a continuous time interval. If we let W denote the time of the occurrence of the first event, then W is a continuous random variable. Theorem 4.3.3 shows that W has an exponential distribution.

Theorem 4.3.3. Consider a Poisson process with parameter λ. Let W denote the time of the occurrence of the first event. W has an exponential distribution with $\beta = 1/\lambda$.

Proof. The distribution function F for W is given by

$$F(w) = P[W \le w] = 1 - P[W > w]$$

The first occurrence of the event will take place after time w only if no occurrences of the event are recorded in the time interval $[0, w]$. Let X denote the number of occurrences of the event in this time interval. X is a Poisson random variable with parameter λw. Thus

$$P[W > w] = P[X = 0] = \frac{e^{-\lambda w}(\lambda w)^0}{0!} = e^{-\lambda w}$$

By substitution we obtain

$$F(w) = 1 - P[W > w] = 1 - e^{-\lambda w}$$

Since in the continuous case the derivative of the cumulative distribution function is the density

$$F'(w) = f(w) = \lambda e^{-\lambda w}$$

This is the density for an exponential random variable with $\beta = 1/\lambda$.

The next example illustrates the use of this theorem.

Example 4.3.2. Some strains of paramecia produce and secrete "killer" particles that will cause the death of a sensitive individual if contact is made. All paramecia unable to produce killer particles are sensitive. The mean number of killer particles emitted by a killer paramecium is 1 every 5 hours. In observing such a paramecium, what is the probability that we must wait at most 4 hours before the first particle is emitted? Considering the measurement unit to be one hour, we are observing a Poisson process with $\lambda = 1/5$. By Theorem 4.3.3, W, the time at which the first killer particle is emitted, has an exponential distribution with $\beta = 1/\lambda = 5$. The density for W is

$$f(w) = (1/5)e^{-w/5} \qquad w > 0$$

The desired probability is given by

$$\begin{aligned} P[W \le 4] &= \int_0^4 (1/5)e^{-w/5} dw \\ &= -e^{-w/5}\Big|_0^4 \\ &= 1 - e^{-4/5} \doteq .5507 \end{aligned}$$

Since an exponential random variable is also a gamma random variable, the average time that we must wait until the first killer particle is emitted is

$$E[W] = \alpha\beta = 1 \cdot 5 = 5 \text{ hours}$$

Chi-Squared Distribution

The gamma distribution gives rise to another important family of random variables, namely, the *chi-squared* family. This distribution is used extensively in applied statistics. Among other things, it provides the basis for making inferences about the variance of a population based on a sample. At this time we consider only the theoretical properties of the chi-squared distribution. You will see many examples of its use in later chapters.

Definition 4.3.3 (Chi-squared distribution). Let X be a gamma random variable with $\beta = 2$ and $\alpha = \gamma/2$ for γ a positive integer. X is said to have a chi-squared distribution with γ degrees of freedom. We denote this variable by X_γ^2.

Note that a chi-squared random variable is completely specified by stating its degrees of freedom. By applying Theorem 4.3.2, we see that the mean of a chi-squared random variable is γ, its degrees of freedom; its variance is 2γ, twice its degrees of freedom. Figure 4.4(*c*) gives the graph of the density of a chi-squared random variable with 4 degrees of freedom.

Since the chi-squared distribution arises so often in practice, extensive tables of its cumulative distribution function have been derived. One such table is Table IV of App. A. In the table, degrees of freedom appear as row headings, probabilities appear as column headings, and points associated with those probabilities are listed in the body of the table. Notationally, we shall use χ_r^2 to denote that point associated with a chi-squared random variable such that

$$P[X_\gamma^2 \geq \chi_r^2] = r$$

That is, χ_r^2 is the point such that the area to its *right* is r. Technically speaking, we should write $\chi_{r,\gamma}^2$ since the value of the point does depend on both the probability desired and the number of degrees of freedom associated with the random variable. However, in applications the value of γ will be obvious. Therefore to simplify notation, we use only a single subscript. The use of this notation is illustrated in the following example.

Example 4.3.3. Consider a chi-squared random variable with 10 degrees of freedom. Find the value of $\chi_{.05}^2$. This point is shown in Fig. 4.5. By definition the area to the right of this point is .05; the area to its left is .95. The column probabilities in Table IV give the area to the *left* of the point listed. Thus to find $\chi_{.05}^2$, we look in row 10 and column .95 and see that $\chi_{.05}^2 = 18.3$.

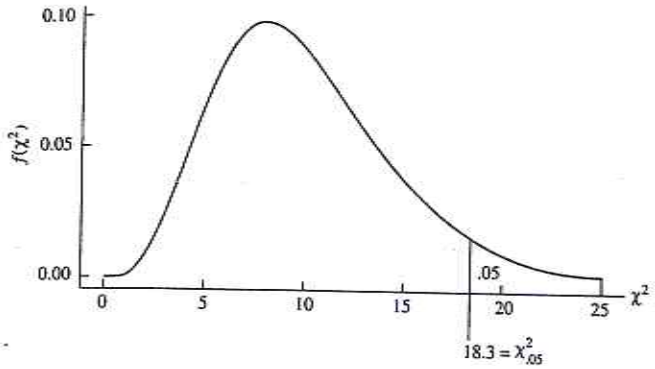


FIGURE 4.5
$P[X_{10}^2 \geq \chi_{.05}^2] = .05$ and $P[X_{10}^2 < \chi_{.05}^2] = .95$.

4.4 NORMAL DISTRIBUTION

The normal distribution is a distribution that underlies many of the statistical methods used in data analysis. It was first described in 1733 by De Moivre as being the limiting form of the binomial density as the number of trials becomes infinite. This discovery did not get much attention, and the distribution was "discovered" again by both Laplace and Gauss a half-century later. Both men dealt with problems of astronomy, and each derived the normal distribution as a distribution that seemingly described the behavior of errors in astronomical measurements. The distribution is often referred to as the "gaussian" distribution.

Definition 4.4.1 (Normal distribution). A random variable X with density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2} \qquad -\infty < x < \infty$$
$$-\infty < \mu < \infty$$
$$\sigma > 0$$

is said to have a normal distribution with parameters μ and σ.

One implication of this definition is that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)[(x-\mu)/\sigma]^2} dx = 1$$

To verify this requires a transformation to polar coordinates. This technique is beyond the mathematical level assumed here. A detailed proof can be found in [49]. Note that Definition 4.4.1 states only that μ is a real number and that σ is positive. As you might suspect from the notation used, the parameters that appear in the

equation for the density for a normal random variable are, in fact, its mean and its standard deviation. This can be verified once we know the moment generating function for X. Our next theorem gives us the form for this important function.

Theorem 4.4.1. Let X be normally distributed with parameters μ and σ. The moment generating function for X is given by

$$m_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

For the proof of this theorem, see Appendix C.

It is now easy to show that the parameters that appear in the definition of the normal density are actually the mean and the standard deviation of the variable.

Theorem 4.4.2. Let X be a normal random variable with parameters μ and σ. Then μ is the mean of X and σ is its standard deviation.

Proof. The moment generating function for X is

$$m_X(t) = e^{\mu t + \sigma^2 t^2/2}$$

and

$$\frac{dm_X(t)}{dt} = e^{\mu t + \sigma^2 t^2/2}(\mu + \sigma^2 t)$$

By Theorem 3.4.2 the mean of X is given by

$$E[X] = \left.\frac{dm_X(t)}{dt}\right|_{t=0} = e^{\mu \cdot 0 + \sigma^2 0^2/2}(\mu + \sigma^2 \cdot 0) = \mu$$

as claimed. The proof of the remainder of the theorem is left as an exercise.

The graph of the density of a normal random variable is a symmetric, bell-shaped curve centered at its mean. The points of inflection occur at $\mu \pm \sigma$.

Example 4.4.1. One of the major contributors to air pollution is hydrocarbons emitted from the exhaust system of automobiles. Let X denote the number of grams of hydrocarbons emitted by an automobile per mile. Assume that X is normally distributed with a mean of 1 gram and a standard deviation of .25 gram. The density for X is given by

$$f(x) = \frac{1}{\sqrt{2\pi}(.25)} e^{-(1/2)[(x-1)/.25]^2}$$

The graph of this density is a symmetric, bell-shaped curve centered at $\mu = 1$ with inflection points at $\mu \pm \sigma$, or $1 \pm .25$. A sketch of the density is given in Fig. 4.6.

One point must be made. Theoretically speaking, a normal random variable must be able to assume any value whatsoever. This is clearly unrealistic here. It is

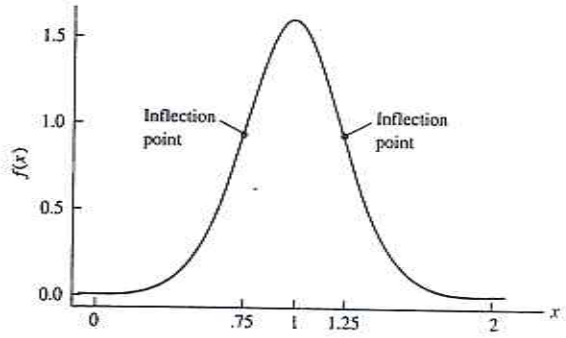


FIGURE 4.6
Graph of the density for a normal random variable with mean 1 and standard deviation .25.

impossible for an automobile to emit a negative amount of hydrocarbons. When we say that X is normally distributed, we mean that over the range of physically reasonable values of X, the given normal curve yields acceptable probabilities. With this understanding, we can at least approximate, for example, the probability that a randomly selected automobile will emit between .9 and 1.54 grams of hydrocarbons by finding the area under the graph of f between these two points.

Standard Normal Distribution

There are infinitely many normal random variables each of which is uniquely characterized by the two parameters μ and σ. To calculate probabilities associated with a specific normal curve requires that one integrate the normal density over a particular interval. However, the normal density is not integrable in closed form. To find areas under the normal curve requires the use of numerical integration techniques. A simple algebraic transformation is employed to overcome this problem. By means of this transformation, called the *standardization procedure*, any question about any normal random variable can be transformed to an equivalent question concerning a normal random variable with mean 0 and standard deviation 1. This particular normal random variable is denoted by Z and is called the *standard normal* variable.

Theorem 4.4.3 (Standardization theorem). Let X be normal with mean μ and standard deviation σ. The variable $(X - \mu)/\sigma$ is standard normal.

You have already verified that the transformation yields a random variable with mean 0 and standard deviation 1 (see Chap. 3, Exercise 21). To prove that the transformed variable is normal requires the use of moment generating function techniques to be introduced in Chap. 7.

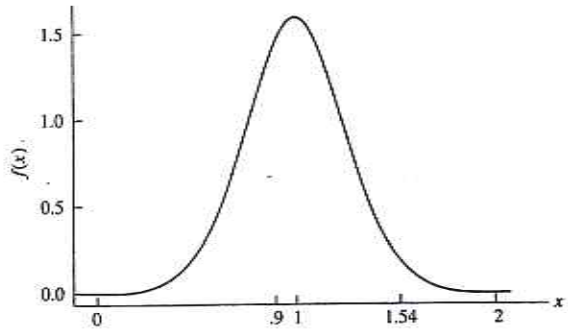


FIGURE 4.7
Shaded area = $P[.9 \le X \le 1.54]$.

The cumulative distribution function for the standard normal random variable is given in Table V of App. A. The use of the standardization theorem and this table is illustrated in the following example.

Example 4.4.2. Let X denote the number of grams of hydrocarbons emitted by an automobile per mile. Assuming that X is normal with $\mu = 1$ gram and $\sigma = .25$ gram, find the probability that a randomly selected automobile will emit between .9 and 1.54 grams of hydrocarbons per mile. The desired probability is shown in Fig. 4.7. To find $P[.9 \le X \le 1.54]$, we first standardize by subtracting the mean of 1 and dividing by the standard deviation of .25 across the inequality. That is,

$$P[.9 \le X \le 1.54] = P[(.9 - 1)/.25 \le (X - 1)/.25 \le (1.54 - 1)/.25]$$

The random variable $(X - 1)/.25$ is now Z. Therefore the problem is to find $P[-.4 \le Z \le 2.16]$ from Table V. We first express the desired probability in terms of the cumulative distribution as follows:

$$\begin{aligned} P[-.4 \le Z \le 2.16] &= P[Z \le 2.16] - P[Z < -.4] \\ &= P[Z \le 2.16] - P[Z \le -.4] \qquad (Z \text{ is continuous}) \\ &= F(2.16) - F(-.4) \end{aligned}$$

$F(2.16)$ is found by locating the first two digits (2.1) in the column headed z; since the third digit is 6, the desired probability of .9846 is found in the row labeled 2.1 and the column labeled .06. Similarly, $F(-.4)$ or .3446 is found in the row labeled -0.4 and the column labeled .00. We now see that the probability that a randomly selected automobile will emit between .9 and 1.54 grams of hydrocarbons per mile is

$$\begin{aligned} P[.9 \le X \le 1.54] &= P[-.4 \le Z \le 2.16] \\ &= F(2.16) - F(-.4) \\ &= .9846 - .3446 = .64 \end{aligned}$$

Interpreting this probability as a percentage, we can say that 64% of the automobiles in operation emit between .9 and 1.54 grams of hydrocarbons per mile driven.

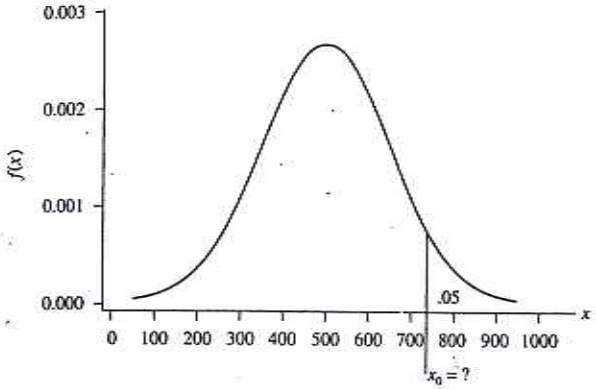


FIGURE 4.8
$P[X \ge x_0] = .05$.

We shall have occasion to read Table V in reverse. That is, given a particular probability r we shall need to find the point with r of the area to its right. This point is denoted by z_r. Thus, notationally, z_r denotes that point associated with a standard normal random variable such that

$$P[Z \ge z_r] = r$$

To see how this need arises, consider Example 4.4.3.

Example 4.4.3. Let X denote the amount of radiation that can be absorbed by an individual before death ensues. Assume that X is normal with a mean of 500 roentgens and a standard deviation of 150 roentgens. Above what dosage level will only 5% of those exposed survive? Here we are asked to find the point x_0 shown in Fig. 4.8. In terms of probabilities, we want to find the point x_0 such that

$$P[X \ge x_0] = .05$$

Standardizing gives

$$\begin{aligned} P[X \ge x_0] &= P\left[\frac{X - 500}{150} \ge \frac{x_0 - 500}{150}\right] \\ &= P\left[Z \ge \frac{x_0 - 500}{150}\right] = .05 \end{aligned}$$

Thus $(x_0 - 500)/150$ is the point on the standard normal curve with 5% of the area under the curve to its right and 95% to its left. That is, $(x_0 - 500)/150$ is the point $z_{.05}$. From Table V the numerical value of this point is approximately 1.645 (we have interpolated). Equating these, we get

$$\frac{x_0 - 500}{150} = 1.645$$

Solving this equation for x_0 gives the desired dosage level:

$$x_0 = 150(1.645) + 500 = 746.75 \text{ roentgens}$$

4.5 NORMAL PROBABILITY RULE AND CHEBYSHEV'S INEQUALITY

It is sometimes useful to have a quick way of determining which values of a random variable are common and which are considered to be rare. In the case of a normally distributed random variable, a rule of thumb, called the *normal probability rule*, can be developed easily. This rule is given in Theorem 4.5.1.

Theorem 4.5.1 (Normal probability rule). Let X be normally distributed with parameters μ and σ. Then

$$P[-\sigma < X - \mu < \sigma] \doteq .68$$
$$P[-2\sigma < X - \mu < 2\sigma] \doteq .95$$
$$P[-3\sigma < X - \mu < 3\sigma] \doteq .997$$

Proof. Note that division by σ yields

$$P[-\sigma < X - \mu < \sigma] = P\left[-1 < \frac{X-\mu}{\sigma} < 1\right]$$

By Theorem 4.4.3, $(X - \mu)/\sigma$ follows the standard normal distribution. From Table V of App. A,

$$P[-1 < Z < 1] = .8413 - .1587 = .6826$$

This probability can be rounded to .68. The other results given in the theorem are proved similarly.

The normal probability rule can be expressed in terms of percentages. In particular, it implies that in repeated sampling from a normal distribution approximately 68% of the observed values of X should lie within 1 standard deviation of its mean; 95% should lie within two standard deviations, and 99.7% within 3 standard deviations of the mean. Thus an observed value that falls farther than 3 standard deviations from μ is indeed rare, since such values occur with probability .003. This rule will be used later to obtain a quick estimate of the standard deviation of a normally distributed random variable.

Figure 4.9 illustrates the normal probability rule as it applies to the standard normal distribution. Recall that for this distribution $\sigma = 1$, $2\sigma = 2$, and $3\sigma = 3$.

Chebyshev's Inequality

A second rule of thumb that can be used to gauge the rarity of observed values of a random variable is *Chebyshev's inequality*. This inequality was derived by the

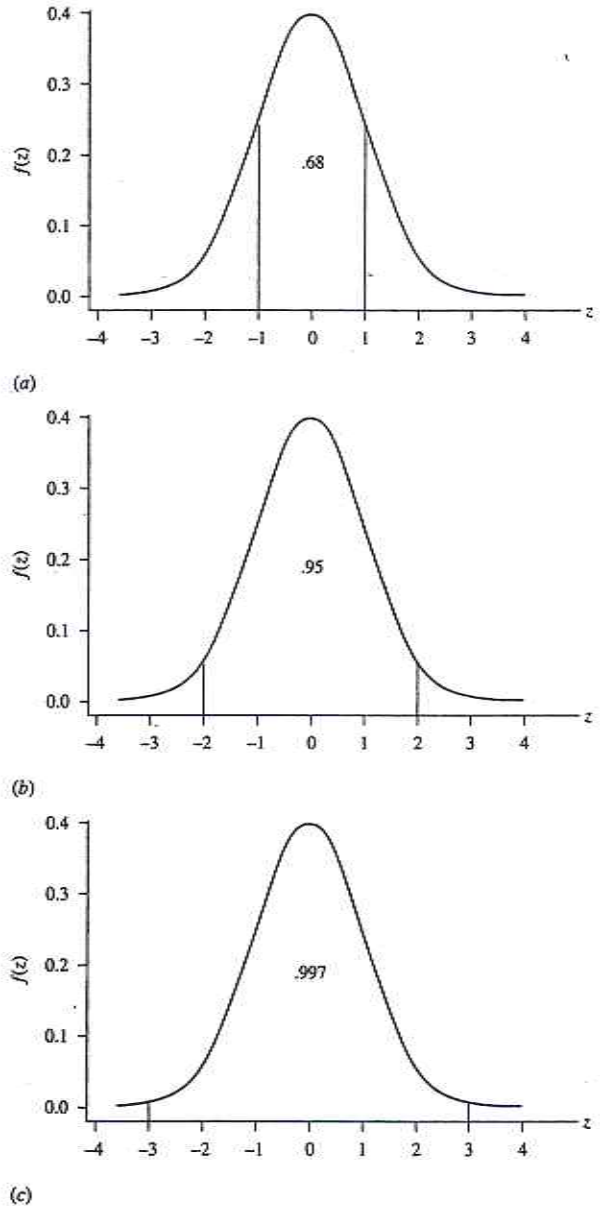


FIGURE 4.9
(*a*) The probability that a normally distributed random variable will lie within one standard deviation of its mean is approximately .68 or 68%.
(*b*) The probability that a normally distributed random variable will lie within two standard deviations of its mean is approximately .95 or 95%.
(*c*) The probability that a normally distributed random variable will lie within three standard deviations of its mean is approximately .997 or 99.7%.

Russian probabilist P. L. Chebyshev (Tchebysheff, 1821–1894). The inequality differs from the normal probability rule in that it does *not* require that the random variable involved be normally distributed. Although we shall prove the theorem in the continuous setting, continuity is *not* required. The inequality holds for any random variable.

Theorem 4.5.2 (Chebyshev's inequality). Let X be a random variable with mean μ and standard deviation σ. Then for any positive number k,

$$P[|X - \mu| < k\sigma] \geq 1 - \frac{1}{k^2}$$

See Appendix C for the proof of this theorem.

Some examples will clarify the difference between Theorems 4.5.1 and 4.5.2.

Example 4.5.1. The viscosity of a fluid can be measured roughly by dropping a small ball into a calibrated tube containing the fluid and observing X, the time that it takes for the ball to drop a measured distance. Assume that this random variable is normally distributed with a mean of 20 s and a standard deviation of .5 s. By the normal probability rule, approximately 95% of the observed values of X will lie within 1 s (2 standard deviations) of the mean. That is, X will fall between 19 and 21 s with probability .95. Since Chebyshev's inequality applies to any random variable, it is appropriate here. This inequality guarantees that X will fall between 19 and 21 s (within $k = 2$ standard deviations of its mean) with probability *at least* $1 - 1/k^2 = .75$. Note that when the random variable in question is normally distributed, the normal probability rule yields a stronger statement than does Chebyshev's inequality.

Example 4.5.2. The safety record of an industrial plant is measured in terms of M, the total staffing-hours worked without a serious accident. Past experience indicates that M has a mean of 2 million with a standard deviation of .1 million. A serious accident has just occurred. Would it be unusual for the next serious accident to occur within the next 1.6 million staffing-hours? To answer this question, we must assess $P[M \leq 1.6]$. Since we have no reason to assume that M is normally distributed, the normal probability rule is inappropriate here. However, we know from Chebyshev's inequality with $k = 4$ that

$$P[1.6 < M < 2.4] \geq 1 - (1/16) = .9375$$

This implies that

$$P[M \leq 1.6] + P[M \geq 2.4] \leq .0625$$

Since it is possible for M to exceed 2.4, we can safely say that

$$P[M \leq 1.6] < .0625$$

No stronger statement can be made without some knowledge of the shape of the density of M. However, if it is known that the density is symmetric, then we can go one step further and state that

$$P[M \leq 1.6] \leq .0625/2 = .03125$$

4.6 NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION

The binomial tables given in this text or in any other text are necessarily limited in scope due to the fact that n can vary from 1 to infinity and p can assume any value between 0 and 1. It is impossible to table every combination of n and p. Due to the advances in computer and calculator technology, it is now possible to find exact binomial probabilities for any combination of n and p. Prior to this time, the normal curve was used to give good approximations of binomial probabilities. The technique introduced in this section is still useful in situations in which the needed technology tools are not readily available. To see how such approximations were suggested, we consider four binomial random variables each with probability of success .4 but with differing values for n. The densities for these variables, obtained from Table I of App. A, together with a sketch for each, are given in Fig. 4.10(a) to (d).

The point to note from these diagrams is made in Fig. 4.10(d). Namely, it is not hard to imagine a smooth bell curve that closely fits the block diagram shown. This suggests that binomial probabilities represented by one or more blocks in the diagram can be approximated reasonably well by a carefully selected area under an appropriately chosen normal curve. Which of the infinitely many normal curves is appropriate? Common sense indicates that the normal variable selected should have the same mean and variance as the binomial variable that it approximates. Theorem 4.6.1 summarizes these ideas.

Theorem 4.6.1 (Normal approximation to the binomial distribution). Let X be binomial with parameters n and p. For large n, X is approximately normal with mean np and variance $np(1 - p)$.

The proof of this theorem is based on the Central Limit Theorem, which will be considered in Chap. 7. Admittedly, Theorem 4.6.1 is a bit vague in the sense that the word "large" is not well defined. In the strictest mathematical sense, "large" means as n approaches infinity. For most practical purposes the approximation is acceptable for values of n and p such that either $p \leq .5$ and $np > 5$ or $p > .5$ and $n(1 - p) > 5$.

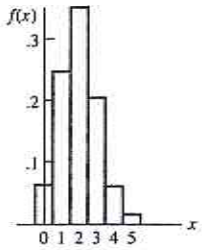
Example 4.6.1. A study is performed to investigate the connection between maternal smoking during pregnancy and birth defects in children. Of the mothers studied, 40% smoke and 60% do not. When the babies were born, 20 were found to have some sort of birth defect. Let X denote the number of children whose mother smoked while pregnant. If there is no relationship between maternal smoking and birth defects, then X is binomial with $n = 20$ and $p = .4$. What is the probability that 12 or more of the affected children had mothers who smoked?

To answer this question, we need to find $P[X \geq 12]$ under the assumption that X is binomial with $n = 20$ and $p = .4$. This probability, .0565, can be found from Table I of App. A. Note that since $p = .4 \leq .5$ and $np = 20(.4) = 8 > 5$, the normal approximation should give a result quite close to .0565. We shall approximate probabilities associated with X using a normal random variable Y with mean $np = 20(.4) = 8$ and standard deviation $\sqrt{np(1-p)} = \sqrt{20(.4)(.6)} = \sqrt{4.8}$.

$n = 5$, $p = .4$

x	$f(x)$
0	.0778
1	.2592
2	.3456
3	.2304
4	.0768
5	.0102

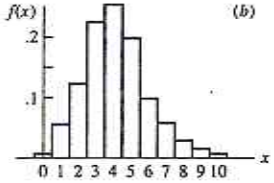
(a)



$n = 10$, $p = .4$

x	$f(x)$
0	.0060
1	.0404
2	.1209
3	.2150
4	.2508
5	.2007
6	.1114
7	.0425
8	.0106
9	.0016
10	.0001

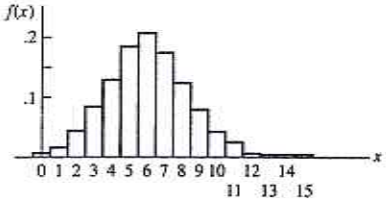
(b)



$n = 15$, $p = .4$

x	$f(x)$	x	$f(x)$
0	.0005	8	.1181
1	.0047	9	.0612
2	.0219	10	.0245
3	.0634	11	.0074
4	.1268	12	.0016
5	.1859	13	.0003
6	.2066	14	~0
7	.1771	15	~0

(c)



$n = 20$, $p = .4$

x	$f(x)$	x	$f(x)$
0	~0	11	.0710
1	.0005	12	.0355
2	.0031	13	.0145
3	.0124	14	.0049
4	.0350	15	.0013
5	.0746	16	.0003
6	.1244	17	~0
7	.1659	18	~0
8	.1797	19	~0
9	.1597	20	~0
10	.1172		

(d)

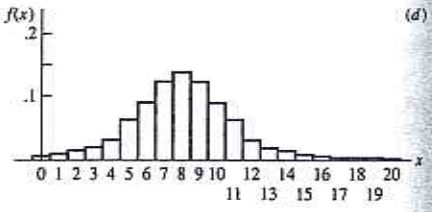


FIGURE 4.10
Density for X binomial: (a) $n = 5, p = .4$; (b) $n = 10, p = .4$; (c) $n = 15, p = .4$; (d) $n = 20, p = .4$.

The exact probability of .0565 is given by the sum of the areas of the blocks centered at 12, 13, 14, 15, 16, 17, 18, 19, and 20, as shown in Fig. 4.11. The approximate probability is given by the area under the normal curve shown above 11.5. That is,

$$P[X \geq 12] \doteq P[Y \geq 11.5]$$

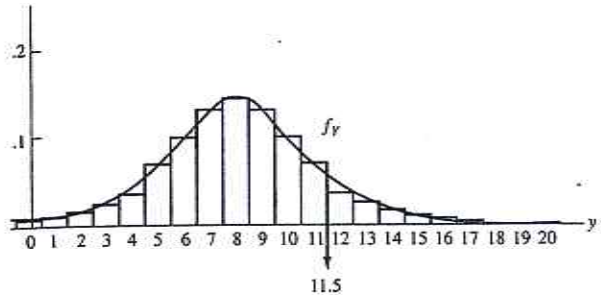


FIGURE 4.11
$P[X \geq 12]$ = area of shaded blocks $\doteq$ area under curve beyond 11.5.

The number .5 is called the *half-unit correction* for continuity. It is subtracted from 12 in the approximation because otherwise half the area of the block centered at 12 will be inadvertently ignored, leading to an unnecessary error in the calculation. From this point on the calculation is routine:

$$\begin{aligned} P[X \geq 12] &\doteq P[Y \geq 11.5] \\ &= P\left[\frac{Y-8}{\sqrt{4.8}} \geq \frac{11.5-8}{\sqrt{4.8}}\right] \\ &= P[Z \geq 1.59] \\ &= 1 - .9441 = .0559 \end{aligned}$$

Note that even with n as small as 20, the approximated value of .0559 compares quite favorably with the exact value of .0565. In practice, of course, one would not approximate a probability that could be found directly from a binomial table. This was done here only for comparative purposes.

4.7 WEIBULL DISTRIBUTION AND RELIABILITY

In 1951 W. Weibull introduced a distribution that has been found to be useful in a variety of physical applications. It arises quite naturally in the study of reliability as we shall show. The most general form for the Weibull density is given by

$$f(x) = \alpha\beta(x-\gamma)^{\beta-1}e^{-\alpha(x-\gamma)^{\beta}} \qquad \begin{matrix} x > \gamma \\ \alpha > 0 \\ \beta > 0 \end{matrix}$$

The implication of this definition of the density is that there is some minimum or "threshold" value γ below which the random variable X cannot fall. In most physical applications this value is 0. For this reason, we shall define the Weibull density with this fact in mind. Be careful when reading scientific literature to note the form of the Weibull density being used.

Definition 4.7.1 (Weibull distribution). A random variable X is said to have a Weibull distribution with parameters α and β if its density is given by

$$f(x) = \alpha\beta x^{\beta-1}e^{-\alpha x^\beta} \qquad \begin{array}{l} x > 0 \\ \alpha > 0 \\ \beta > 0 \end{array}$$

It is easy to verify that the function given in Definition 4.7.1 is a density. (See Exercise 61.) We shall find the mean of this distribution directly rather than by means of the moment generating function.

Theorem 4.7.1. Let X be a Weibull random variable with parameters α and β. The mean and variance of X are given by

$$\mu = \alpha^{-1/\beta}\,\Gamma(1 + 1/\beta)$$

and

$$\sigma^2 = \alpha^{-2/\beta}\,\Gamma(1 + 2/\beta) - \mu^2$$

Proof. By Definition 4.2.1,

$$\begin{aligned} E[X] &= \int_0^\infty x\alpha\beta x^{\beta-1}e^{-\alpha x^\beta}dx \\ &= \int_0^\infty \alpha\beta x^\beta e^{-\alpha x^\beta}dx \end{aligned}$$

Let $z = \alpha x^\beta$. This implies that

$$x = (z/\alpha)^{1/\beta} \qquad \text{and} \qquad dx = (1/\alpha\beta)(z/\alpha)^{1/\beta-1}\,dz$$

By substitution, it is seen that

$$\begin{aligned} E[X] &= \int_0^\infty \alpha\beta(z/\alpha)e^{-z}(1/\alpha\beta)(z/\alpha)^{1/\beta-1}dz \\ &= \int_0^\infty (z/\alpha)^{1/\beta}e^{-z}dz \\ &= \alpha^{-1/\beta}\int_0^\infty z^{1/\beta}e^{-z}dz \end{aligned}$$

The integral on the right is, by definition, $\Gamma(1 + 1/\beta)$. (See Definition 4.3.1.) Thus we have shown that the mean of the Weibull distribution is

$$\mu = E[X] = \alpha^{-1/\beta}\Gamma(1 + 1/\beta)$$

as claimed. The remainder of the proof is outlined as an exercise. (See Exercises 62 and 63.)

The graph of the Weibull density varies depending on the values of α and β. The general shape resembles that of the gamma density with the curve becoming more symmetric as the value of β increases.

Example 4.7.1. Let X be a Weibull random variable with $\beta = 1$. The density for X is

$$f(x) = \alpha e^{-\alpha x} \qquad \begin{array}{l} x > 0 \\ \alpha > 0 \end{array}$$

Note that this is the density for an *exponential* random variable. That is, the exponential distribution is a special case of the Weibull distribution with $\beta = 1$. By Theorem 4.7.1

$$\begin{aligned} \mu &= \alpha^{-1/\beta}\Gamma(1 + 1/\beta) = (1/\alpha)\Gamma(2) = 1/\alpha \cdot 1! \equiv 1/\alpha \\ \sigma^2 &= \alpha^{-2/\beta}\Gamma(1 + 2/\beta) - \mu^2 \\ &= 1/\alpha^2\Gamma(3) - (1/\alpha)^2 \\ &= 2/\alpha^2 - 1/\alpha^2 = 1/\alpha^2 \end{aligned}$$

Note that these results are consistent with those obtained by viewing this random variable as being exponential. (See Exercise 33.)

Reliability

As we have said, the Weibull distribution frequently arises in the study of reliability. Reliability studies are concerned with assessing whether or not a system functions adequately under the conditions for which it was designed. Interest centers on describing the behavior of the random variable X, the time to failure of a system that cannot be repaired once it fails to operate. Three functions come into play when assessing reliability. These are the failure density f, the reliability function R, and ρ, the failure or hazard rate of the distribution. To understand how these functions are defined, consider some system being put into operation at time $t = 0$. We observe the system until it eventually fails. Let X denote the time of the failure. This random variable is continuous and a priori can assume any value in the interval $(0, \infty)$. The density f, for X, is called the *failure density* for the component. The *reliability function*, R, is defined to be the probability that the component will not fail before time t. Thus

$$\begin{aligned} R(t) &= 1 - P[\text{component will fail before time } t] \\ &= 1 - \int_0^t f(x)\,dx \\ &= 1 - F(t) \end{aligned}$$

where F is the cumulative distribution function for X. To define ρ, the hazard rate function, consider a time interval $[t, t + \Delta t]$ of length Δt. We define the force of mortality or hazard rate function over this interval by

TABLE 4.1
Continuous distributions

Name	Density		Moment generating function		Mean	Variance
Gamma	$\frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1}e^{-x/\beta}$	$\alpha > 0$ $\beta > 0$ $x > 0$	$(1-\beta t)^{-\alpha}$		$\alpha\beta$	$\alpha\beta^2$
Exponential	$\frac{1}{\beta} e^{-x/\beta}$	$x > 0$ $\beta > 0$	$(1-\beta t)^{-1}$		β	β^2
Chi-squared	$\frac{1}{\Gamma(\gamma/2)2^{\gamma/2}} x^{\gamma/2-1}e^{-x/2}$	$x > 0$ γ a positive integer	$(1-2t)^{-\gamma/2}$		γ	2γ
Uniform	$\frac{1}{b-a}$	$a < x < b$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$ 1	$t \neq 0$ $t = 0$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Cauchy	$\frac{1}{\pi}\frac{a}{a^2+(x-b)^2}$	$-\infty < x < \infty$ $-\infty < b < \infty$ $a > 0$	Does not exist		Does not exist	Does not exist
Normal	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-1/2\left(\frac{x-\mu}{\sigma}\right)^2\right]$	$-\infty < x < \infty$ $\sigma > 0$ $-\infty < \mu < \infty$	$e^{\mu t+\sigma^2 t^2/2}$		μ	σ^2
Weibull	$\alpha\beta x^{\beta-1}e^{-\alpha x^{\beta}}$	$x > 0$ $\alpha > 0$ $\beta > 0$			$\alpha^{-1/\beta}\Gamma\left(1+\frac{1}{\beta}\right)$	$\alpha^{-2/\beta}\Gamma\left(1+\frac{2}{\beta}\right)-\mu^2$

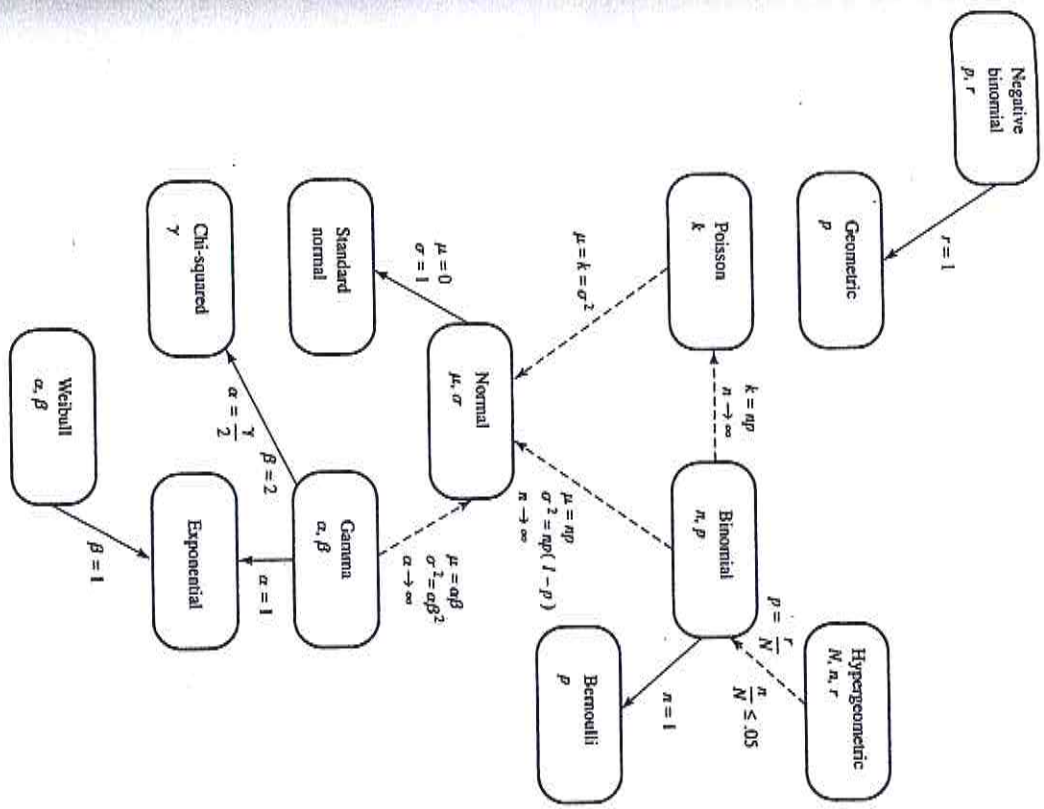


FIGURE 4.16
Some interrelationships among common distributions.

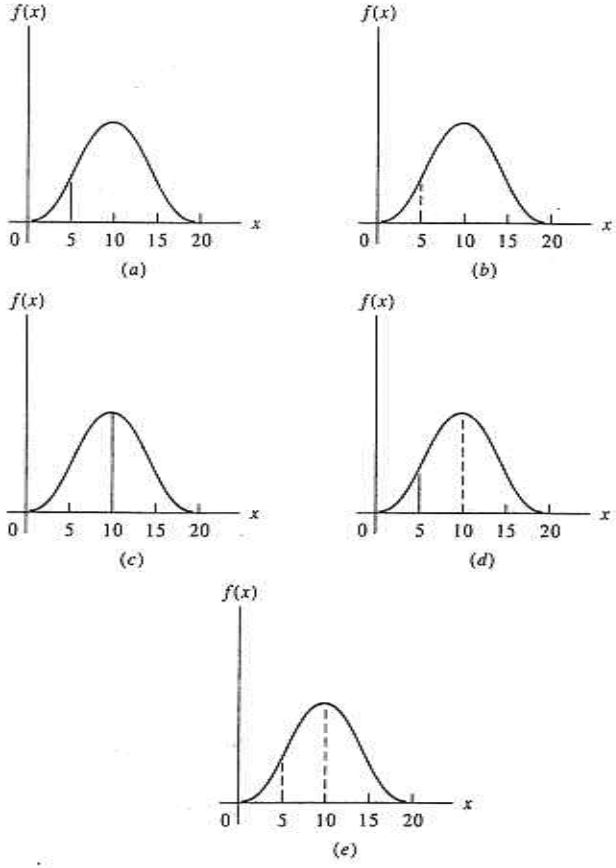


FIGURE 14.17

EXERCISES

Section 4.1

1. Consider the function

$$f(x) = kx \qquad 2 \le x \le 4$$

(*a*) Find the value of k that makes this a density for a continuous random variable.
(*b*) Find $P[2.5 \le X \le 3]$.
(*c*) Find $P[X = 2.5]$.
(*d*) Find $P[2.5 < X \le 3]$.

2. Consider the areas shown in Fig. 4.17. In each case, state what probability is being depicted. What is the relationship between the areas depicted in Figs. 4.17(*a*) and (*b*)? Between those in Figs. 4.17(*d*) and (*e*)?
3. Let X denote the length in minutes of a long-distance telephone conversation. Assume that the density for X is given by

$$f(x) = (1/10)e^{-x/10} \qquad x > 0$$

(*a*) Verify that f is a density for a continuous random variable.
(*b*) Assuming that f adequately describes the behavior of the random variable X, find the probability that a randomly selected call will last at most 7 minutes; at least 7 minutes; exactly 7 minutes.
(*c*) Would it be unusual for a call to last between 1 and 2 minutes? Explain, based on the probability of this occurring.
(*d*) Sketch the graph of f and indicate in the sketch the area corresponding to each of the probabilities found in part (*b*).
4. Some plastics in scrapped cars can be stripped out and broken down to recover the chemical components. The greatest success has been in processing the flexible polyurethane cushioning found in these cars. Let X denote the amount of this material, in pounds, found per car. Assume that the density for X is given by

$$f(x) = \frac{1}{\ln 2}\frac{1}{x} \qquad 25 \le x \le 50$$

(*a*) Verify that f is a density for a continuous random variable.
(*b*) Use f to find the probability that a randomly selected auto will contain between 30 and 40 pounds of polyurethane cushioning.
(*c*) Sketch the graph of f, and indicate in the sketch the area corresponding to the probability found in part (*b*).
5. *(Continuous uniform distribution.)* A random variable X is said to be uniformly distributed over an interval (a, b) if its density is given by

$$f(x) = \frac{1}{b-a} \qquad a < x < b$$

(*a*) Show that this is a density for a continuous random variable.
(*b*) Sketch the graph of the uniform density.
(*c*) Shade the area in the graph of part (*b*) that represents $P[X \le (a+b)/2]$.
(*d*) Find the probability pictured in part (*c*).
(*e*) Let (c, d) and (e, f) be subintervals of (a, b) of equal length. What is the relationship between $P[c \le X \le d]$ and $P[e \le X \le f]$? Generalize the idea suggested by this example, thus justifying the name "uniform" distribution.
6. If a pair of coils were placed around a homing pigeon and a magnetic field was applied that reverses the earth's field, it is thought that the bird would become disoriented. Under these circumstances it is just as likely to fly in one direction as in any other. Let θ denote the direction in radians of the bird's initial flight. See Fig. 4.18. θ is uniformly distributed over the interval $[0, 2\pi]$.
(*a*) Find the density for θ.

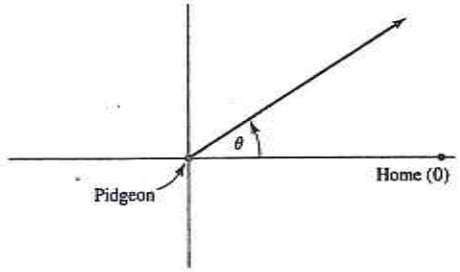


FIGURE 4.18
θ = direction of the initial flight of a homing pigeon measured in radians.

(*b*) Sketch the graph of the density. The uniform distribution is sometimes called the "rectangular" distribution. Do you see why?
(*c*) Shade the area corresponding to the probability that a bird will orient within $\pi/4$ radians of home, and find this area using plane geometry.
(*d*) Find the probability that a bird will orient within $\pi/4$ radians of home by integrating the density over the appropriate region(s), and compare your answer to that obtained in part (*c*).
(*e*) If 10 birds are released independently and at least seven orient within $\pi/4$ radians of home, would you suspect that perhaps the coils are not disorienting the birds to the extent expected? Explain, based on the probability of this occurring.

7. Use Definition 4.1.2 to show that for a continuous random variable X, $P[X = a] = 0$ for every real number a. *Hint:* Write $P[X = a]$ as $P[a \le X \le a]$.

8. Express each of the probabilities depicted in Fig. 4.16 in terms of the cumulative distribution function F.

9. Consider the random variable of Exercise 1.
(*a*) Find the cumulative distribution function F.
(*b*) Use F to find $P[2.5 \le X \le 3]$, and compare your answer to that obtained previously.
(*c*) Find $F'(x)$, and verify that your result is the density given in Exercise 1.

10. *(Uniform distribution.)* Find the general expression for the cumulative distribution function for a random variable X that is uniformly distributed over the interval (a, b). See Exercise 5.

11. *(Uniform distribution.)* Consider the random variable of Exercise 6.
(*a*) Use Exercise 10 to find the cumulative distribution function F.
(*b*) Find $F'(x)$, and verify that your result is, as expected, the uniform density over the interval $[0, 2\pi]$.

12. Find the cumulative distribution function for the random variable of Exercise 3. Use F to find $P[1 \le X \le 2]$, and compare your answer to that obtained previously.

13. Find the cumulative distribution function for the random variable of Exercise 4. Use F to find $P[30 \le X \le 40]$, and compare your answer to that obtained previously.

14. In parts (*a*) and (*b*) proposed cumulative distributions are given. In each case, find the "density" that would be associated with each, and decide whether it really does define a valid continuous density. If it does not, explain what property fails.
(*a*) Consider the function F defined by

$$F(x) = \begin{cases} 0 & x < -1 \\ x + 1 & -1 \le x \le 0 \\ 1 & x > 0 \end{cases}$$

(*b*) Consider the function defined by

$$F(x) = \begin{cases} 0 & x \le 0 \\ x^2 & 0 < x \le 1/2 \\ (1/2)x & 1/2 < x \le 1 \\ 1 & x > 1 \end{cases}$$

Section 4.2

15. Consider the random variable X with density

$$f(x) = (1/6)x \qquad 2 \le x \le 4$$

(*a*) Find $E[X]$.
(*b*) Find $E[X^2]$.
(*c*) Find σ^2 and σ.

16. Let X denote the amount in pounds of polyurethane cushioning found in a car. (See Exercise 4.) The density for X is given by

$$f(x) = \frac{1}{\ln 2}\frac{1}{x} \qquad 25 \le x \le 50$$

Find the mean, variance, and standard deviation for X.

17. Let X denote the length in minutes of a long-distance telephone conversation. The density for X is given by

$$f(x) = (1/10)e^{-x/10} \qquad x > 0$$

(*a*) Find the moment generating function, $m_X(t)$.
(*b*) Use $m_X(t)$ to find the average length of such a call.
(*c*) Find the variance and standard deviation for X.

18. *(Uniform distribution.)* The density for a random variable X distributed uniformly over (a, b) is

$$f(x) = \frac{1}{b - a} \qquad a < x < b$$

Use Definition 4.2.1 to show that

$$E[X] = \frac{a + b}{2} \quad \text{and} \quad \text{Var } X = \frac{(b - a)^2}{12}$$

CHAPTER 5

JOINT DISTRIBUTIONS

Thus far interest has centered on a single random variable of either the discrete or the continuous type. Such random variables are called *univariate*. Problems do arise in which two random variables are to be studied simultaneously. For example, we might wish to study the yield of a chemical reaction in conjunction with the temperature at which the reaction is run. Typical questions to ask are: "Is the yield independent of the temperature?" or, "What is the average yield if the temperature is 40° C?" To answer questions of this type, we need to study what are called *two-dimensional* or *bivariate random variables* of both the discrete and continuous type. In this chapter we present a brief introduction to the basic theoretical concepts underlying these variables. These concepts form the basis for the study of regression analysis and correlation, topics of extreme importance in applied statistics. (See Chaps. 11 and 12.)

5.1 JOINT DENSITIES AND INDEPENDENCE

We begin by considering two-dimensional random variables and their density functions. The definitions presented here are natural extensions of those presented for a single random variable in Chaps. 3 and 4. (See Definition 3.2.1 and 4.1.2.)

Definition 5.1.1 (Discrete joint density). Let X and Y be discrete random variables. The ordered pair (X, Y) is called a two-dimensional discrete random variable. A function f_{XY} such that

$$f_{XY}(x, y) = P[X = x \text{ and } Y = y]$$

is called the joint density for (X, Y).

Again, let us point out that in the discrete case some statisticians prefer to use the term "probability function" or "probability mass function" rather than the term "density." We shall use the term "density" and the notation f_{XY} in both the discrete and the continuous cases for consistency of notation and terminology.

Note that the purpose of the density here is the same as in the past—to allow us to compute the probability that the random variable (X, Y) will assume specific values. As in the one-dimensional case, f_{XY} is nonnegative since it represents a probability. Furthermore, if the density is summed over all possible values of X and Y, it must sum to 1. That is, the necessary and sufficient conditions for a function to be a joint density for a two-dimensional discrete random variable are as follows:

Necessary and Sufficient Conditions for a Function to Be a Discrete Joint Density

1. $f_{XY}(x, y) \geq 0$
2. $\sum_{\text{all } x} \sum_{\text{all } y} f_{XY}(x, y) = 1$

The joint density in the discrete case is sometimes expressed in closed form. However, it is more common to present the density in table form.

Example 5.1.1. In an automobile plant two tasks are performed by robots. The first entails welding two joints; the second, tightening three bolts. Let X denote the number of defective welds and Y the number of improperly tightened bolts produced per car. Since X and Y are each discrete, (X, Y) is a two-dimensional discrete random variable. Past data indicates that the joint density for (X, Y) is as shown in Table 5.1. Note that each entry in the table is a number between 0 and 1 and therefore can be interpreted as a probability. Furthermore,

$$\sum_{x=0}^{2} \sum_{y=0}^{3} f_{XY}(x, y) = .840 + .030 + .020 + \cdots + .001 = 1$$

as required. The probability that there will be no errors made by the robots is given by

$$P[X = 0 \text{ and } Y = 0] = f_{XY}(0, 0) = .840$$

The probability that there will be exactly one error made is

TABLE 5.1

x/y	0	1	2	3
0	.840	.030	.020	.010
1	.060	.010	.008	.002
2	.010	.005	.004	.001

$$P[X = 1 \text{ and } Y = 0] + P[X = 0 \text{ and } Y = 1] = f_{XY}(1, 0) + f_{XY}(0, 1)$$
$$= .060 + .030$$
$$= .09$$

The probability that there will be no improperly tightened bolts is $P[Y = 0]$. Note that this probability, which concerns only the random variable Y, can be obtained by summing $f_{XY}(x, 0)$ over all values of X. That is,

$$P[Y = 0] = \sum_{x=0}^{2} f_{XY}(x, 0)$$
$$= P[X = 0 \text{ and } Y = 0] + P[X = 1 \text{ and } Y = 0] + P[X = 2 \text{ and } Y = 0]$$
$$= .840 + .060 + .010 = .91$$

Marginal Distributions: Discrete

Given the joint density for a two-dimensional discrete random variable (X, Y), it is easy to derive the individual densities for X and Y. The manner in which this is done is suggested by the method used to answer the last question posed in Example 5.1.1. To find the density for Y alone, we sum the joint density over all values of X; to find the density for X alone, we sum over Y. When the joint density is given in table form, it is customary to report the individual densities for X and Y in the margins of the joint density table. For this reason, the densities for X and Y alone are called *marginal densities*. This idea is formalized in Definition 5.1.2.

Definition 5.1.2 (Discrete marginal densities). Let (X, Y) be a two-dimensional discrete random variable with joint density f_{XY}. The marginal density for X, denoted by f_X, is given by

$$f_X(x) = \sum_{\text{all } y} f_{XY}(x, y)$$

The marginal density for Y, denoted by f_Y, is given by

$$f_Y(y) = \sum_{\text{all } x} f_{XY}(x, y)$$

Example 5.1.2. Table 5.2 gives the joint density for the random variable (X, Y) of Example 5.1.1. It also displays the marginal densities for X, the number of defective welds, and Y, the number of improperly tightened bolts per car. Note that the marginal density for X is obtained by summing across the rows of the table; that for Y is obtained by summing down the columns.

Joint and Marginal Distributions: Continuous

The idea of a two-dimensional continuous random variable and continuous joint density can be developed by extending Definition 4.1.1 to more than one variable.

TABLE 5.2

x/y	0	1	2	3	$f_X(x)$
0	.840	.030	.020	.010	.900
1	.060	.010	.008	.002	.080
2	.010	.005	.004	.001	.020
$f_Y(y)$	.910	.045	.032	.013	1.000

Definition 5.1.3 (Continuous joint density). Let X and Y be continuous random variables. The ordered pair (X, Y) is called a two-dimensional continuous random variable. A function f_{XY} such that

1. $f_{XY}(x, y) \geq 0 \qquad -\infty < x < \infty \qquad -\infty < y < \infty$
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x, y)\, dy\, dx = 1$
3. $P[a \leq X \leq b \text{ and } c \leq Y \leq d] = \int_a^b \int_c^d f_{XY}(x, y)\, dy\, dx$

for a, b, c, d real is called the joint density for (X, Y).

Even though the joint density is defined for all real values x and y, we shall follow the convention of specifying its equation only over those regions for which it may be nonzero. Recall that in the case of a single continuous random variable, probabilities correspond to areas. In the case of a two-dimensional continuous random variable, probabilities correspond to *volumes*. These ideas are illustrated in Example 5.1.3.

Example 5.1.3. In a healthy individual age 20 to 29 years, the calcium level in the blood, X, is usually between 8.5 and 10.5 milligrams per deciliter (mg/dl) and the cholesterol level, Y, is usually between 120 and 240 mg/dl. Assume that for a healthy individual in this age group the random variable (X, Y) is uniformly distributed over the rectangle whose corners are (8.5, 120), (8.5, 240), (10.5, 120), (10.5, 240). That is, assume that the joint density for (X, Y) is

$$f_{XY}(x, y) = c \qquad 8.5 \leq x \leq 10.5 \qquad 120 \leq y \leq 240$$

To be a density, c must be chosen so that

$$\int_{8.5}^{10.5}\int_{120}^{240} c\, dy\, dx = 1$$

That is, c must be chosen so that the volume of the rectangular solid shown in Fig. 5.1(*a*) is 1. To find c, we can use geometry or complete the indicated integration as shown below.

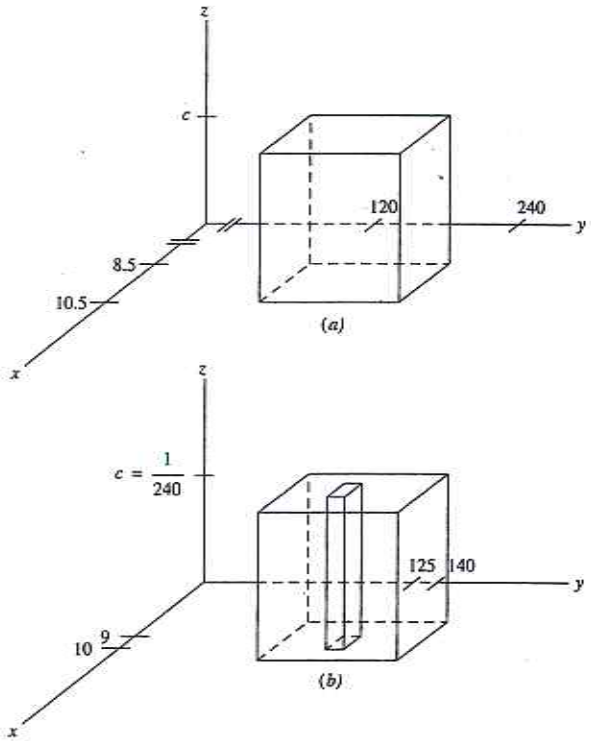


FIGURE 5.1
(*a*) Volume of the solid whose base is a rectangle with corners (8.5, 120), (8.5, 240), (10.5, 120), and (10.5, 240) and height c is 1; (*b*) $P[9 \le X \le 10 \text{ and } 125 \le Y \le 140]$ = volume of solid whose base is a rectangle with corners (9, 125), (9, 140), (10, 125), (10, 140) and height $c = 1/240$.

$$\int_{8.5}^{10.5}\int_{120}^{240} c\,dy\,dx = 1$$

$$c\int_{8.5}^{10.5}(240 - 120)\,dx = 1$$

$$120c(10.5 - 8.5) = 1$$

$$240c = 1$$

$$c = 1/240$$

Let us now use the joint density to find the probability that an individual's calcium level will lie between 9 and 10 mg/dl, whereas the cholesterol level is between 125 and 140 mg/dl. This probability corresponds to the volume of the solid shown in Fig. 5.1(*b*). This probability is

$$P[9 \le X \le 10 \text{ and } 125 \le Y \le 140] = \int_9^{10}\int_{125}^{140} 1/240\,dy\,dx$$

$$= 1/240\int_9^{10}(140 - 125)\,dx$$

$$= 15/240$$

To define "marginal" densities in the continuous case, we replace summation by integration. This yields the following definition.

Definition 5.1.4 (Continuous marginal densities). Let (X, Y) be a two-dimensional continuous random variable with joint density f_{XY}. The marginal density for X, denoted by f_X, is given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$$

The marginal density for Y, denoted by f_Y, is given by

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx$$

We illustrate the idea of marginal densities in Examples 5.1.4 and 5.1.5.

Example 5.1.4. Let X denote an individual's blood calcium level and Y his or her blood cholesterol level. The joint density for (X, Y) is

$$f_{XY}(x, y) = 1/240 \qquad 8.5 \le x \le 10.5$$
$$120 \le y \le 240$$

The marginal densities for X and Y are

$$f_X(x) = \int_{120}^{240} 1/240\,dy = 1/2 \qquad 8.5 \le x \le 10.5$$

$$f_Y(y) = \int_{8.5}^{10.5} 1/240\,dx = 2/240 \qquad 120 \le y \le 240$$

To find the probability that a healthy individual has a cholesterol level between 150 and 200, we can use either the joint density or the marginal density for Y. That is,

$$P[150 \le Y \le 200] = \int_{8.5}^{10.5}\int_{150}^{200} 1/240\,dy\,dx = 100/240$$

or

$$P[150 \le Y \le 200] = \int_{150}^{200} 2/240\,dy = 100/240$$

Note that both X and Y are uniformly distributed.

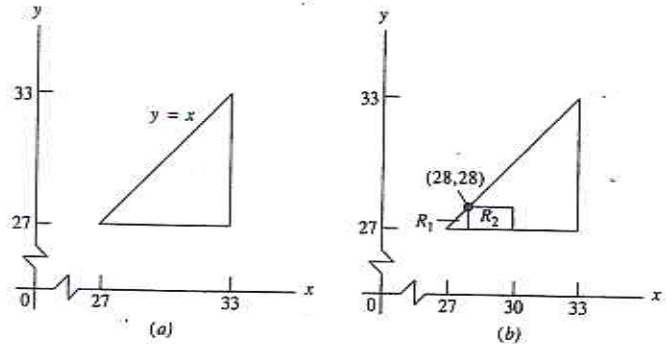


FIGURE 5.2
(*a*) The joint density $f(x, y) = c/x$ is defined over the triangular region bounded by $y = 27$, $y = x$, and $x = 33$.
(*b*)

$$P[X \le 30 \text{ and } Y \le 28] = \iint_{R_1} c/x \, dy \, dx + \iint_{R_2} c/x \, dy \, dx$$
$$= \int_{27}^{28}\int_{27}^{x} c/x \, dy \, dx + \int_{28}^{30}\int_{27}^{28} c/x \, dy \, dx$$

or

$$P[X \le 30 \text{ and } Y \le 28] = \int_{27}^{28}\int_{y}^{30} c/x \, dx \, dy.$$

Example 5.1.5. In studying the behavior of air support roofs, the random variables X, the inside barometric pressure (in inches of mercury), and Y, the outside pressure, are considered. Assume that the joint density for (X, Y) is given by

$$f_{XY}(x, y) = c/x \qquad 27 \le y \le x \le 33$$
$$c = 1/(6 - 27 \ln 33/27) \doteq 1.72$$

The region in the plane over which this joint density is defined is shown in Fig. 5.2(*a*). The marginal densities for X and Y are given by

$$f_X(x) = \int_{27}^{x} c/x \, dy = (c/x)y\Big|_{27}^{x} = c(1 - 27/x) \qquad 27 \le x \le 33$$

$$f_Y(y) = \int_{y}^{33} c/x \, dx = c(\ln 33 - \ln y) \qquad 27 \le y \le 33$$

Let us find the probability that the inside pressure is at most 30 and the outside pressure is at most 28. That is, let us find $P[X \le 30 \text{ and } Y \le 28]$. The region over which the joint density is to be integrated is shown in Fig. 5.2(*b*). Integration can be done with respect to y and then x or vice versa. In the former case the problem must be split into two pieces, since the boundaries for y change at the point (28, 28). In the latter case integration can be accomplished more easily. The integrals required in the two cases are

Case I:

$$P[X \le 30 \text{ and } Y \le 28] = \int_{27}^{28}\int_{27}^{x} c/x \, dy \, dx + \int_{28}^{30}\int_{27}^{28} c/x \, dy \, dx$$

Case II:

$$P[X \le 30 \text{ and } Y \le 28] = \int_{27}^{28}\int_{y}^{30} c/x \, dx \, dy$$

Since case II requires less effort, we find $P[X \le 30 \text{ and } Y \le 28]$ as follows:

$$P[X \le 30 \text{ and } Y \le 28] = \int_{27}^{28}\int_{y}^{30} c/x \, dx \, dy$$
$$= c\int_{27}^{28} [\ln 30 - \ln y] dy$$
$$= c\left[y \ln 30\Big|_{27}^{28} - \int_{27}^{28} \ln y \, dy\right]$$
$$= c\left[\ln 30 - (y \ln y - y)\Big|_{27}^{28}\right]$$
$$= c[\ln 30 - 28 \ln 28 + 27 \ln 27 + 1]$$
$$\doteq c(.09) = 1.72(.09) = .15$$

It is left as an exercise to show that the same result is obtained via case I. (See Exercise 6.)

Independence

There is one other point to be made in this section. Recall that two events are independent if knowledge of the fact that one has occurred gives us no clue as to the likelihood that the other will occur. Suppose that X and Y are discrete random variables such that knowledge of the value assumed by one gives us no clue as to the value assumed by the other. We would like to think of these random variables as being "independent" and would like a mathematical characterization of this property. The characterization is suggested by the following argument. Let X and Y be discrete. Let A_1 denote the event that $X = x$, and let A_2 denote the event that $Y = y$. If X and Y are independent in the intuitive sense, then A_1 and A_2 are independent events. By Definition 2.3.1

$$P[A_1 \cap A_2] = P[A_1]P[A_2]$$

Substituting, we see that

$$P[X = x \text{ and } Y = y] = P[X = x]P[Y = y]$$

or

$$f_{XY}(x, y) = f_X(x)f_Y(y)$$

It seems that, at least in the discrete case, independence implies that the *joint density can be expressed as the product of the marginal densities.* This idea provides the

basis for the definition of the term "independent random variables" in both the discrete and continuous cases.

Definition 5.1.5 (Independent random variables). Let X and Y be random variables with joint density f_{XY} and marginal densities f_X and f_Y, respectively. X and Y are independent if and only if

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

for all x and y.

Example 5.1.6

(*a*) The random variables X, the number of defective welds, and Y, the number of improperly tightened bolts per car of Examples 5.1.1 and 5.1.2, are not independent. To verify this, note that from Table 5.2

$$f_{XY}(0, 0) = .84 \neq .9(.91) = .819 = f_X(0) f_Y(0)$$

(*b*) The random variables X, an individual's blood calcium level, and Y, his or her blood cholesterol level as described in Examples 5.1.3 and 5.1.4, are independent. To verify this, note that

$$f_{XY}(x, y) = 1/240 = 1/2 \cdot 2/240 = f_X(x) f_Y(y)$$

An important point should be made here. The assumption that (X, Y) is uniformly distributed leads to the conclusion that X and Y are independent. If this conclusion is *medically unsound*, then another more realistic density should be sought to describe the behavior of the two-dimensional random variable (X, Y).

(*c*) The random variables X and Y, the inside and outside pressure, respectively, on an air support roof of Example 5.1.5 are not independent. This is seen by noting that

$$f_{XY}(x, y) = c/x \neq c(1 - 27/x)c(\ln 33 - \ln y) = f_X(x) f_Y(y)$$

The assumption of nonindependence here is realistic from a physical point of view.

The exercises for Sec. 5.1 provide some practice in dealing with these theoretical ideas. You will see their relationship to data analysis in chapters to come.

5.2 EXPECTATION AND COVARIANCE

In this section we introduce the idea of *expectation* in the case of a two-dimensional random variable. We also study a specific expectation, called the *covariance*, that is useful in describing the behavior of one variable relative to another.

We begin by extending Definitions 3.3.1 and 4.2.1 to the two-dimensional case.

Definition 5.2.1 (Expected value). Let (X, Y) be a two-dimensional random variable with joint density f_{XY}. Let $H(X, Y)$ be a random variable. The expected value of $H(X, Y)$, denoted by $E[H(X, Y)]$ is given by

1. $E[H(X, Y)] = \sum_{\text{all } x} \sum_{\text{all } y} H(x, y) f_{XY}(x, y)$

   provided $\sum_{\text{all } x} \sum_{\text{all } y} |H(x, y)| f_{XY}(x, y)$ exists for (X, Y) discrete;

2. $E[H(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y) f_{XY}(x, y) dy\, dx$

   provided $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |H(x, y)| f_{XY}(x, y) dy\, dx$ exists for (X, Y) continuous.

As in the case of one-dimensional random variables, some functions of X and Y are of more interest than others. In particular, if the joint density for (X, Y) is known, then the average value of X and of Y can be found easily. These are determined as follows:

Univariate Averages Found Via the Joint Density

$$E[X] = \sum_{\text{all } x} \sum_{\text{all } y} x f_{XY}(x, y) \qquad \text{for } (X, Y) \text{ discrete}$$

$$E[Y] = \sum_{\text{all } x} \sum_{\text{all } y} y f_{XY}(x, y)$$

$$E[X] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) dx\, dy \qquad \text{for } (X, Y) \text{ continuous}$$

$$E[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{XY}(x, y) dx\, dy$$

Examples 5.2.1 and 5.2.2 illustrate the use of this definition.

Example 5.2.1. The joint density for the random variable (X, Y) of Example 5.1.1 is given in Table 5.3. X denotes the number of defective welds and Y, the number of improperly tightened bolts produced per car by assembly line robots. Let us use Definition 5.2.1 to find $E[X]$, $E[Y]$, $E[X + Y]$, and $E[XY]$.

$$\begin{aligned} E[X] &= \sum_{x=0}^{2} \sum_{y=0}^{3} x f_{XY}(x, y) \\ &= 0(.840) + 0(.030) + 0(.020) + 0(.010) + 1(.060) + \cdots + 2(.001) \\ &= .12 \end{aligned}$$

TABLE 5.3

x/y	0	1	2	3	$f_X(x)$
0	.840	.030	.020	.010	.900
1	.060	.010	.008	.002	.080
2	.010	.005	.004	.001	.020
$f_Y(y)$	.910	.045	.032	.013	1.000

$$E[Y] = \sum_{x=0}^{2}\sum_{y=0}^{3} yf_{XY}(x, y)$$
$$= 0(.840) + 1(.030) + 2(.020) + 3(.010) + 0(.060) + \cdots + 3(.001)$$
$$= .148$$

$$E[X + Y] = \sum_{x=0}^{2}\sum_{y=0}^{3} (x + y)f_{XY}(x, y)$$
$$= (0 + 0)(.840) + (0 + 1)(.030) + (0 + 2)(.020) + \cdots + (2 + 3)(.001)$$
$$= .268$$

$$E[XY] = \sum_{x=0}^{2}\sum_{y=0}^{3} xyf_{XY}(x, y)$$
$$= (0 \cdot 0)(.840) + (0 \cdot 1)(.030) + (0 \cdot 2)(.020) + \cdots + (2 \cdot 3)(.001)$$
$$= .064$$

There are two points to be made. First, both $E[X]$ and $E[Y]$ were found via the joint density and Definition 5.2.1. These expectations could have been found just as easily from the marginal densities and Definition 3.3.1. (See Exercise 18.) Second, note that $E[X + Y] = E[X] + E[Y]$. This result is consistent with the rules of expectation given in Theorem 3.3.1.

Example 5.2.2. The joint density for the random variable (X, Y), where X denotes the calcium level and Y denotes the cholesterol level in the blood of a healthy individual, is given by

$$f_{XY}(x, y) = 1/240 \qquad 8.5 \le x \le 10.5$$
$$120 \le y \le 240$$

For these variables,

$$E[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf_{XY}(x, y)dy\,dx$$
$$= \int_{8.5}^{10.5}\int_{120}^{240} x(1/240)dy\,dx$$
$$= \int_{8.5}^{10.5} (1/2)x\,dx = x^2/4\Big|_{8.5}^{10.5} = 9.5 \text{ mg/dl}$$

$$E[Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf_{XY}(x, y)\,dy\,dx$$
$$= \int_{8.5}^{10.5}\int_{120}^{240} y(1/240)\,dy\,dx$$
$$= 1/240\int_{8.5}^{10.5} y^2/2\Big|_{120}^{240} dx$$
$$= 1/240\int_{8.5}^{10.5} 21{,}600\,dx = 180 \text{ mg/dl}$$

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x, y)\,dy\,dx$$
$$= \int_{8.5}^{10.5}\int_{120}^{240} xy(1/240)\,dy\,dx$$
$$= 1/240\int_{8.5}^{10.5} xy^2/2\Big|_{120}^{240} dx$$
$$= 1/240\int_{8.5}^{10.5} 21{,}600x\,dx$$
$$= (21{,}600/240)(x^2/2)\Big|_{8.5}^{10.5} = 1710$$

Covariance

Occasionally the expected value of a function of X and Y is of interest in its own right. For instance, in Example 5.2.1, $E[X + Y]$ gives the theoretical average number of errors made by the robots overall. However, we shall be concerned primarily with those expectations that are needed to compute the covariance between X and Y. This term is defined as follows:

Definition 5.2.2 (Covariance). Let X and Y be random variables with means μ_X and μ_Y respectively. The covariance between X and Y, denoted by $\text{Cov}(X, Y)$ or σ_{XY} is given by

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Note that if small values of X tend to be associated with small values of Y and large values of X with large values of Y, then $X - \mu_X$ and $Y - \mu_Y$ will usually have the same algebraic signs. This implies that $(X - \mu_X)(Y - \mu_Y)$ will be positive, yielding a positive covariance. If the reverse is true and small values of X tend to be associated with large values of Y and vice versa, then $X - \mu_X$ and $Y - \mu_Y$ will usually have opposite algebraic signs. This results in a negative value for $(X - \mu_X)(Y - \mu_Y)$, yielding a negative covariance. In this sense covariance is an indication of how X and Y vary relative to one another.

Covariance is seldom computed from Definition 5.2.2. Rather, we apply the following computational formula whose derivation is left as an exercise. (See Exercise 24.)

Theorem 5.2.1 (Computational formula for covariance)

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$$

We illustrate the use of Theorem 5.2.1 by finding the covariance for the random variables of Examples 5.2.1 and 5.2.2.

Example 5.2.3

(*a*) The covariance between X, the number of defective welds, and Y, the number of improperly tightened bolts of Example 5.2.1, is given by

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= .064 - (.12)(.148) = .046\end{aligned}$$

Since $\text{Cov}(X, Y) > 0$, there is a tendency for large values of X to be associated with large values of Y and vice versa. That is, a car with an above average number of defective welds tends also to have an above average number of improperly tightened bolts and vice versa.

(*b*) The covariance between X, an individual's blood calcium level, and Y, his or her blood cholesterol level, has covariance given by

$$\begin{aligned}\text{Cov}(X, Y) &= E[XY] - E[X]E[Y] \\ &= 1710 - (9.5)(180) = 0\end{aligned}$$

A covariance of 0 implies that knowledge that X assumes a value above its mean gives us no indication as the value of Y relative to its mean.

The fact that the covariance between X and Y is 0 in Example 5.2.2 is not a coincidence. It is, of course, due to the fact that $E[XY] = E[X]E[Y]$. It can be shown that this property will hold whenever the random variables X and Y are independent, as they are in Example 5.2.2. This important result is formalized in the following theorem:

Theorem 5.2.2. Let (X, Y) be a two-dimensional random variable with joint density f_{XY}. If X and Y are independent then

$$E[XY] = E[X]E[Y]$$

Proof. We shall prove this theorem in the continuous case. The proof in the discrete case is similar. Assume that (X, Y) has joint density f_{XY} and that X and Y are independent. Let f_X and f_Y denote the marginal densities for X and Y, respectively. By Definition 5.2.1,

TABLE 5.4

x/y	-2	-1	1	2	$f_1(x)$
1	0	1/4	1/4	0	1/2
4	1/4	0	0	1/4	1/2
$f_2(y)$	1/4	1/4	1/4	1/4	1

$$\begin{aligned}E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x, y)dy\,dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dy\,dx \qquad (X \text{ and } Y \text{ are independent}) \\ &= \int_{-\infty}^{\infty} xf_X(x)\int_{-\infty}^{\infty} yf_Y(y)dy\,dx \\ &= \int_{-\infty}^{\infty} xf_X(x)E[Y]dx \\ &= E[Y]\int_{-\infty}^{\infty} xf_X(x)dx = E[Y]E[X]\end{aligned}$$

An immediate consequence of this theorem is the result that we have already noted and observed relative to Example 5.2.2. In particular, *if X and Y are independent, then* $\text{Cov}(X, Y) = 0$. Unfortunately, the converse of this statement is not true. That is, we *cannot* conclude that a zero covariance implies independence. The next example verifies this contention.

Example 5.2.4. The joint density for (X, Y) is given in Table 5.4, from which we see that $E[X] = 5/2$, $E[Y] = 0$, and $E[XY] = 0$, yielding a covariance of 0. It is also easy to see that X and Y are *not* independent. The value assumed by Y does have an effect on that assumed by X. In fact, $X = Y^2$. The value of Y completely determines the value of X!

Covariance gives us only a very rough idea of the relationship between X and Y. We are concerned only with its algebraic sign and not with its magnitude. However, covariance is used to define another measure of the relationship between X and Y which is easier to interpret. This measure, called the *correlation*, is discussed in the next section.

5.3 CORRELATION

Recall that the covariance between X and Y gives only a rough indication of any association that may exist between X and Y. No attempt is made to describe the type or strength of the association. Often it is of interest to know whether or not two random variables are *linearly* related. One measure used to determine this is the Pearson coefficient of correlation, ρ. In this section we define this theoretical measure of linearity; in Chap. 11 we shall discuss how to estimate its value from a data set.

Definition 5.3.1 (Pearson coefficient of correlation). Let X and Y be random variables with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively. The correlation, ρ_{XY}, between X and Y is given by

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{(\text{Var } X)(\text{Var } Y)}}$$

Since we already know how to calculate each of the terms appearing in the above definition, calculating ρ_{XY} (or ρ) from the joint density for (X, Y) is easy. The question is, "How do we interpret ρ once we know its numerical value?" To interpret ρ, we must know its range of possible values. The next theorem shows that, unlike the covariance which can assume any real value, the correlation coefficient is bounded.

Theorem 5.3.1. The correlation coefficient ρ_{XY} for any two random variables X and Y lies between -1 and 1 inclusive.

The proof of this theorem is found in Appendix C.

The next theorem indicates how ρ measures linearity. The point of the theorem is twofold. First, if there is a linear relationship between X and Y, then this fact is reflected in a correlation coefficient of 1 or -1. Second, if $\rho = 1$ or -1, then a linear relationship exists between X and Y. The formal statement of this result is given in Theorem 5.3.2.

Theorem 5.3.2. Let X and Y be random variables with correlation coefficient ρ_{XY}. Then $|\rho_{XY}| = 1$ if and only if $Y = \beta_0 + \beta_1 X$ for some real numbers β_0 and $\beta_1 \neq 0$.

See Appendix C for the proof of this theorem.

If $\rho = 1$, then we say that X and Y have *perfect positive* correlation. Perfect positive correlation implies that $Y = \beta_0 + \beta_1 X$, where $\beta_1 > 0$. This in turn implies that small values of X are associated with small values of Y, and large values of X with large values of Y. Perfect negative correlation implies that $Y = \beta_0 + \beta_1 X$, where $\beta_1 < 0$. Practically speaking, this means that small values of X are associated with large values of Y and vice versa. Unfortunately, random variables seldom assume the easily interpretable values of 1 or -1. However, values of ρ near 1 or -1 do occur and indicate a linear trend. That is, they indicate that, even though no single straight line passes through the points of positive probability, there is a straight line passing through the graph with the property that most of the probability is associated with points lying on or near this straight line. It is equally important to realize what Theorem 5.3.2 is not saying. If $\rho = 0$, we say that X and Y are uncorrelated, but we are *not* saying that they are unrelated. We are saying that if a relationship exists, then it is *not linear*. These ideas are illustrated in Fig. 5.3.

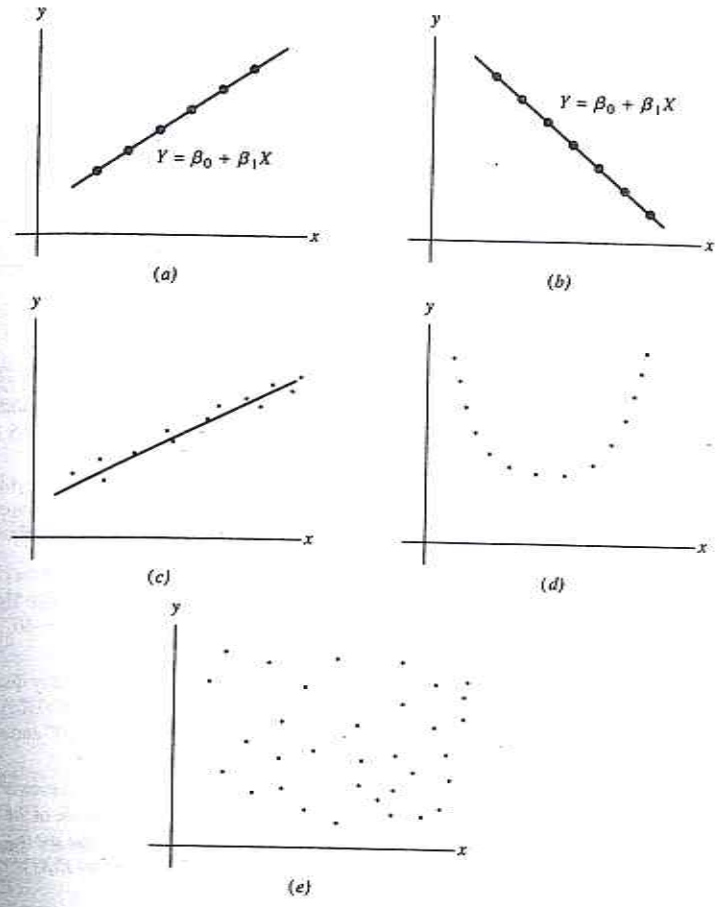


FIGURE 5.3
(*a*) Perfect positive correlation: $\rho = 1, \beta_1 > 0$, all points lie on a straight line with positive slope; (*b*) perfect negative correlation: $\rho = -1, \beta_1 < 0$, all points lie on a straight line with negative slope; (*c*) ρ near 1, points exhibit a linear trend; (*d*) uncorrelated: $\rho = 0$, points indicate a relationship between X and Y, but the relationship is not linear; (*e*) uncorrelated: $\rho = 0$, points are randomly scattered.

Example 5.3.1. To find the correlation between X, the number of defective welds, and Y, the number of improperly tightened bolts produced per car by assembly line robots, we use Table 5.3 to compute $E[X^2]$ and $E[Y^2]$. For these variables

$$E[X^2] = 0^2(.90) + 1^2(.08) + 2^2(.02) = .16$$
$$E[Y^2] = 0^2(.910) + 1^2(.045) + 2^2(.032) + 3^2(.013) = .29$$

In Example 5.2.1, we found that $E[X] = .12$ and $E[Y] = .148$. Therefore

$$\text{Var } X = E[X^2] - (E[X])^2 = .16 - (.12)^2 \doteq .146$$
$$\text{Var } Y = E[Y^2] - (E[Y])^2 = .29 - (.148)^2 \doteq .268$$

In Example 5.2.3 we found that $\text{Cov}(X, Y) = .046$. By Definition 5.3.1,

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var } X \text{ Var } Y}} = \frac{.046}{\sqrt{(.146)(.268)}} \doteq .23$$

Since this value does not appear to lie close to 1, we would not expect the observed values of X and Y to exhibit a strong linear trend.

Exercise 36 points out the relationship between correlation and independence.

5.4 CONDITIONAL DENSITIES AND REGRESSION

In this section we consider two topics that are closely related. These are *conditional densities* and *regression*. To see what is to be done, let us reconsider Example 5.1.5.

Example 5.4.1. In Example 5.1.5 we considered the random variable (X, Y) where X is the inside and Y the outside barometric pressure on an air support roof. Suppose we are interested in studying the inside pressure when the outside pressure is fixed at $y = 30$. There are three important points to understand:

1. The inside pressure will vary even though the outside pressure is constant. Therefore it makes sense to talk about "the random variable X given that $y = 30$." We shall denote this new random variable by $X|y = 30$.
2. Since $X|y = 30$ is a random variable in its own right, it has a probability distribution. Therefore it makes sense to ask, "What is the density for $X|y = 30$?" We shall call this density the "conditional density for X given that $y = 30$" and shall denote it by $f_{X|y=30}$.
3. Since the inside pressure varies even though the outside pressure is constant, it makes sense to ask, "What is the mean or average pressure on the inside of the roof when the outside pressure is 30?" That is, we can ask, "What is the mean value for the random variable $X|y = 30$?" This mean value is denoted by $E[X|y = 30]$ or $\mu_{X|y=30}$.

In general, the conditional density for X given $Y = y$, denoted by $f_{X|y}$, is a function that allows us to find the probability that X assumes specific values based on knowledge of the value assumed by the random variable Y. To see how to define $f_{X|y}$ let us assume that (X, Y) is discrete with joint density f_{XY} and marginal densities f_X and f_Y. Let A_1 denote the event that $X = x$ and A_2 denote the event that $Y = y$. From Definition 2.2.1,

$$P[A_1|A_2] = \frac{P[A_1 \cap A_2]}{P[A_2]}$$

Substituting, we see that

$$P[X = x|Y = y] = \frac{P[X = x \text{ and } Y = y]}{P[Y = y]} = \frac{f_{XY}(x, y)}{f_Y(y)}$$

In the discrete case the conditional density for X given $Y = y$ is the ratio of the joint density for (X, Y) to the marginal density for Y. This observation provides the motivation for the definition of the term "conditional density" in both the discrete and continuous cases. In the formal definition, note that the roles of X and Y can be reversed.

Definition 5.4.1 (Conditional density). Let (X, Y) be a two-dimensional random variable with joint density f_{XY} and marginal densities f_X and f_Y. Then

1. The conditional density for X given $Y = y$, denoted by $f_{X|y}$, is given by
$$f_{X|y} = \frac{f_{XY}(x, y)}{f_Y(y)} \qquad f_Y(y) > 0$$
2. The conditional density for Y given $X = x$, denoted by $f_{Y|x}$, is given by
$$f_{Y|x}(y) = \frac{f_{XY}(x, y)}{f_X(x)} \qquad f_X(x) > 0$$

The use of this definition is illustrated in Example 5.4.2.

Example 5.4.2. The joint density for the random variable (X, Y), where X is the inside and Y is the outside pressure on an air support roof, is given by

$$f_{XY}(x, y) = c/x \qquad 27 \le y \le x \le 33$$
$$c = 1/(6 - 27 \ln 33/27)$$

From Example 5.1.5 the marginal densities for X and Y are

$$f_X(x) = c(1 - 27/x) \qquad 27 \le x \le 33$$

and

$$f_Y(y) = c(\ln 33 - \ln y) \qquad 27 \le y \le 33$$

The conditional density for X given $Y = y$ is

$$f_{X|y}(x) = \frac{f_{XY}(x, y)}{f_Y(y)}$$
$$= \frac{c/x}{c(\ln 33 - \ln y)} = \frac{1}{x(\ln 33 - \ln y)} \qquad y \le x \le 33$$

To find the probability that the inside pressure exceeds 32 given that the outside pressure is 30, we let $y = 30$ in the above expression. We then integrate the conditional density over values of X that exceed 32. That is,

$$P[X > 32|y = 30] = \int_{32}^{33} \frac{1}{x(\ln 33 - \ln 30)}\, dx$$
$$= \frac{\ln x}{\ln 33 - \ln 30}\Big|_{32}^{33}$$
$$= \frac{\ln 33 - \ln 32}{\ln 33 - \ln 30} \doteq .32$$