$$f(y) = \begin{cases} \frac{1}{25}y & 0 \le y < 5 \\ \frac{2}{5} - \frac{1}{25}y & 5 \le y \le 10 \\ 0 & y < 0 \text{ or } y > 10 \end{cases}$$

a. Sketch a graph of the pdf of Y.
b. Verify that $\int_{-\infty}^{\infty} f(y)\,dy = 1$.
c. What is the probability that total waiting time is at most 3 min?
d. What is the probability that total waiting time is at most 8 min?
e. What is the probability that total waiting time is between 3 and 8 min?
f. What is the probability that total waiting time is either less than 2 min or more than 6 min?

Consider again the pdf of X = time headway given in Example 4.5. What is the probability that time headway is

a. At most 6 sec?
b. More than 6 sec? At least 6 sec?
c. Between 5 and 6 sec?

10. A family of pdf's that has been used to approximate the distribution of income, city population size, and size of firms is the Pareto family. The family has two parameters, k and θ, both > 0, and the pdf is

$$f(x; k, \theta) = \begin{cases} \frac{k \cdot \theta^k}{x^{k+1}} & x \ge \theta \\ 0 & x < \theta \end{cases}$$

a. Sketch the graph of $f(x; k, \theta)$.
b. Verify that the total area under the graph equals 1.
c. If the rv X has pdf $f(x; k, \theta)$, for any fixed $b > \theta$, obtain an expression for $P(X \le b)$.
d. For $\theta < a < b$, obtain an expression for the probability $P(a \le X \le b)$.

4.2 Cumulative Distribution Functions and Expected Values

Several of the most important concepts introduced in the study of discrete distributions also play an important role for continuous distributions. Definitions analogous to those in Chapter 3 involve replacing summation by integration.

The Cumulative Distribution Function

The cumulative distribution function (cdf) $F(x)$ for a discrete rv X gives, for any specified number x, the probability $P(X \le x)$. It is obtained by summing the pmf $p(y)$ over all possible values y satisfying $y \le x$. The cdf of a continuous rv gives the same probabilities $P(X \le x)$ and is obtained by integrating the pdf $f(y)$ between the limits $-\infty$ and x.

DEFINITION

The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

For each x, $F(x)$ is the area under the density curve to the left of x. This is illustrated in Figure 4.5, where $F(x)$ increases smoothly as x increases.

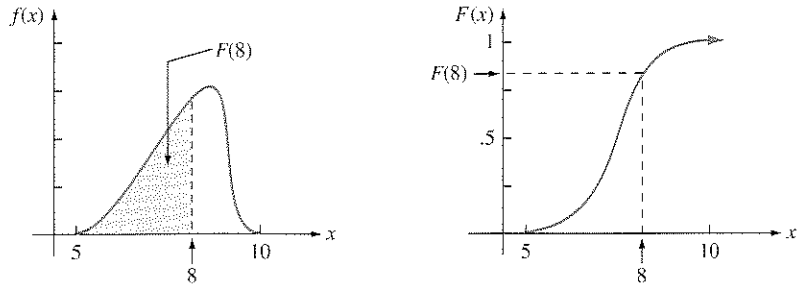

Figure 4.5 A pdf and associated cdf

Example 4.6 Let X, the thickness of a certain metal sheet, have a uniform distribution on $[A, B]$. The density function is shown in Figure 4.6. For $x < A$, $F(x) = 0$, since there is no area under the graph of the density function to the left of such an x. For $x \geq B$, $F(x) = 1$, since all the area is accumulated to the left of such an x. Finally, for $A \leq x \leq B$,

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_{A}^{x} \frac{1}{B-A}\,dy = \frac{1}{B-A} \cdot y \Big|_{y=A}^{y=x} = \frac{x-A}{B-A}$$

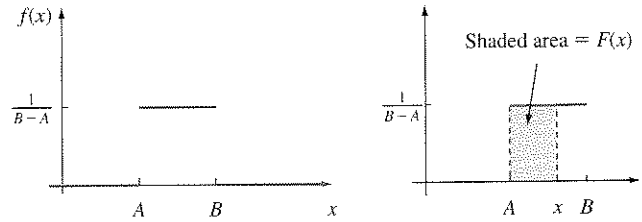

Figure 4.6 The pdf for a uniform distribution

The entire cdf is

$$F(x) = \begin{cases} 0 & x < A \\ \dfrac{x-A}{B-A} & A \leq x < B \\ 1 & x \geq B \end{cases}$$

The graph of this cdf appears in Figure 4.7.

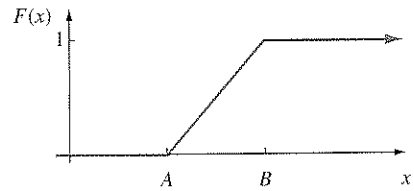

Figure 4.7 The cdf for a uniform distribution

Using $F(x)$ to Compute Probabilities

The importance of the cdf here, just as for discrete rv's, is that probabilities of various intervals can be computed from a formula for or table of $F(x)$.

PROPOSITION

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \leq X \leq b) = F(b) - F(a)$$

Figure 4.8 illustrates the second part of this proposition; the desired probability is the shaded area under the density curve between a and b, and it equals the difference between the two shaded cumulative areas. This is different from what is appropriate for a discrete integer valued random variable (e.g., binomial or Poisson): $P(a \le X \le b) = F(b) - F(a-1)$ when a and b are integers.

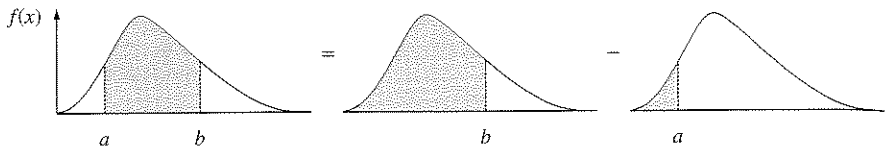

Figure 4.8 Computing $P(a \le X \le b)$ from cumulative probabilities

Example 4.7 Suppose the pdf of the magnitude X of a dynamic load on a bridge (in newtons) is given by

$$f(x) = \begin{cases} \frac{1}{8} + \frac{3}{8}x & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

For any number x between 0 and 2,

$$F(x) = \int_{-\infty}^{x} f(y)\,dy = \int_0^x \left(\frac{1}{8} + \frac{3}{8}y\right) dy = \frac{x}{8} + \frac{3}{16}x^2$$

Thus

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{8} + \frac{3}{16}x^2 & 0 \le x \le 2 \\ 1 & 2 < x \end{cases}$$

The graphs of $f(x)$ and $F(x)$ are shown in Figure 4.9. The probability that the load is between 1 and 1.5 is

$$\begin{aligned} P(1 \le X \le 1.5) &= F(1.5) - F(1) \\ &= \left[\frac{1}{8}(1.5) + \frac{3}{16}(1.5)^2\right] - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] \\ &= \frac{19}{64} = .297 \end{aligned}$$

The probability that the load exceeds 1 is

$$\begin{aligned} P(X > 1) &= 1 - P(X \le 1) = 1 - F(1) = 1 - \left[\frac{1}{8}(1) + \frac{3}{16}(1)^2\right] \\ &= \frac{11}{16} = .688 \end{aligned}$$

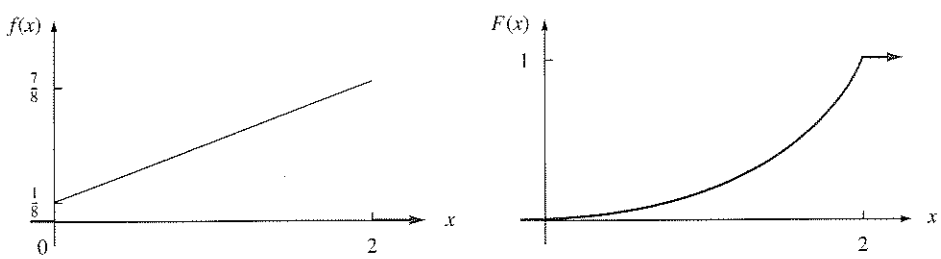

Figure 4.9 The pdf and cdf for Example 4.7

Once the cdf has been obtained, any probability involving X can easily be calculated without any further integration.

Obtaining $f(x)$ from $F(x)$

For X discrete, the pmf is obtained from the cdf by taking the difference between two $F(x)$ values. The continuous analog of a difference is a derivative. The following result is a consequence of the Fundamental Theorem of Calculus.

PROPOSITION

If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists, $F'(x) = f(x)$.

Example 4.8 (Example 4.6 continued)

When X has a uniform distribution, $F(x)$ is differentiable except at $x = A$ and $x = B$, where the graph of $F(x)$ has sharp corners. Since $F(x) = 0$ for $x < A$ and $F(x) = 1$ for $x > B$, $F'(x) = 0 = f(x)$ for such x. For $A < x < B$,

$$F'(x) = \frac{d}{dx}\left(\frac{x - A}{B - A}\right) = \frac{1}{B - A} = f(x)$$

Percentiles of a Continuous Distribution

When we say that an individual's test score was at the 85th percentile of the population, we mean that 85% of all population scores were below that score and 15% were above. Similarly, the 40th percentile is the score that exceeds 40% of all scores and is exceeded by 60% of all scores.

DEFINITION

Let p be a number between 0 and 1. The **(100p)th percentile** of the distribution of a continuous rv X, denoted by $\eta(p)$, is defined by

$$p = F(\eta(p)) = \int_{-\infty}^{\eta(p)} f(y)\, dy \tag{4.2}$$

According to Expression (4.2), $\eta(p)$ is that value on the measurement axis such that $100p\%$ of the area under the graph of $f(x)$ lies to the left of $\eta(p)$ and $100(1 - p)\%$ lies to the right. Thus $\eta(.75)$, the 75th percentile, is such that the area under the graph of $f(x)$ to the left of $\eta(.75)$ is .75. Figure 4.10 illustrates the definition.

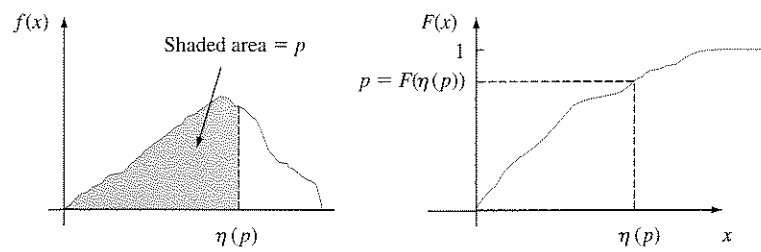

Figure 4.10 The (100p)th percentile of a continuous distribution

Example 4.9 The distribution of the amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous rv X with pdf

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

The cdf of sales for any x between 0 and 1 is

$$F(x) = \int_0^x \frac{3}{2}(1 - y^2)\,dy = \frac{3}{2}\left(y - \frac{y^3}{3}\right)\Bigg|_{y=0}^{y=x} = \frac{3}{2}\left(x - \frac{x^3}{3}\right)$$

The graphs of both $f(x)$ and $F(x)$ appear in Figure 4.11. The $(100p)$th percentile of this distribution satisfies the equation

$$p = F(\eta(p)) = \frac{3}{2}\left[\eta(p) - \frac{(\eta(p))^3}{3}\right]$$

that is,

$$(\eta(p))^3 - 3\eta(p) + 2p = 0$$

For the 50th percentile, $p = .5$, and the equation to be solved is $\eta^3 - 3\eta + 1 = 0$; the solution is $\eta = \eta(.5) = .347$. If the distribution remains the same from week to week, then in the long run 50% of all weeks will result in sales of less than .347 ton and 50% in more than .347 ton.

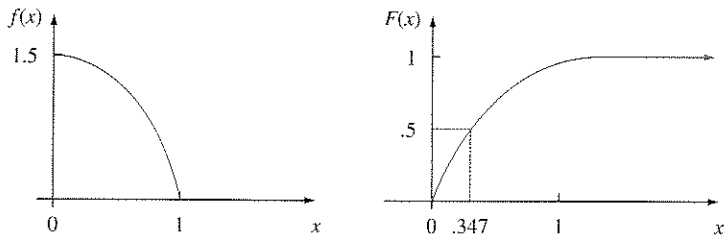

Figure 4.11 The pdf and cdf for Example 4.9

DEFINITION

> The **median** of a continuous distribution, denoted by $\tilde{\mu}$, is the 50th percentile, so $\tilde{\mu}$ satisfies $.5 = F(\tilde{\mu})$. That is, half the area under the density curve is to the left of $\tilde{\mu}$ and half is to the right of $\tilde{\mu}$.

A continuous distribution whose pdf is **symmetric**—the graph of the pdf to the left of some point is a mirror image of the graph to the right of that point—has median $\tilde{\mu}$ equal to the point of symmetry, since half the area under the curve lies to either side of this point. Figure 4.12 gives several examples. The error in a measurement of a physical quantity is often assumed to have a symmetric distribution.

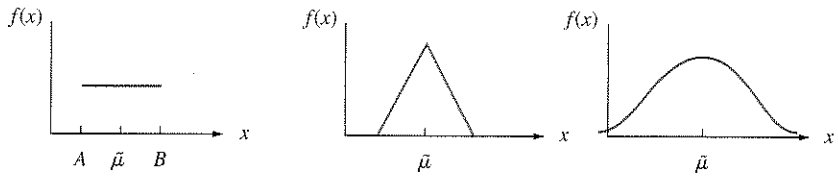

Figure 4.12 Medians of symmetric distributions

Expected Values

For a discrete random variable X, $E(X)$ was obtained by summing $x \cdot p(x)$ over possible X values. Here we replace summation by integration and the pmf by the pdf to get a continuous weighted average.

DEFINITION

The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Example 4.10 (Example 4.9 continued)

The pdf of weekly gravel sales X was

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

so

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \int_0^1 x \cdot \frac{3}{2}(1 - x^2)\, dx$$

$$= \frac{3}{2}\int_0^1 (x - x^3)\, dx = \frac{3}{2}\left(\frac{x^2}{2} - \frac{x^4}{4}\right)\Bigg|_{x=0}^{x=1} = \frac{3}{8}$$

■

When the pdf $f(x)$ specifies a model for the distribution of values in a numerical population, then μ is the population mean, which is the most frequently used measure of population location or center.

Often we wish to compute the expected value of some function $h(X)$ of the rv X. If we think of $h(X)$ as a new rv Y, techniques from mathematical statistics can be used to derive the pdf of Y, and $E(Y)$ can be computed from the definition. Fortunately, as in the discrete case, there is an easier way to compute $E[h(X)]$.

PROPOSITION

If X is a continuous rv with pdf $f(x)$ and $h(X)$ is any function of X, then

$$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x) \cdot f(x)\, dx$$

Example 4.11

Two species are competing in a region for control of a limited amount of a certain resource. Let X = the proportion of the resource controlled by species 1 and suppose X has pdf

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

which is a uniform distribution on [0, 1]. (In her book *Ecological Diversity*, E. C. Pielou calls this the "broken-stick" model for resource allocation, since it is analogous

to breaking a stick at a randomly chosen point.) Then the species that controls the majority of this resource controls the amount

$$h(X) = \max(X, 1 - X) = \begin{cases} 1 - X & \text{if } 0 \le X < \frac{1}{2} \\ X & \text{if } \frac{1}{2} \le X \le 1 \end{cases}$$

The expected amount controlled by the species having majority control is then

$$E[h(X)] = \int_{-\infty}^{\infty} \max(x, 1 - x) \cdot f(x)\, dx = \int_0^1 \max(x, 1 - x) \cdot 1\, dx$$
$$= \int_0^{1/2} (1 - x) \cdot 1\, dx + \int_{1/2}^1 x \cdot 1\, dx = \frac{3}{4}$$

■

For $h(X)$ a linear function, $E[h(X)] = E(aX + b) = aE(X) + b$.

In the discrete case, the variance of X was defined as the expected squared deviation from μ and was calculated by summation. Here again integration replaces summation.

DEFINITION

The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)\, dx = E[(X - \mu)^2]$$

The **standard deviation** (SD) of X is $\sigma_X = \sqrt{V(X)}$.

The variance and standard deviation give quantitative measures of how much spread there is in the distribution or population of x values. The easiest way to compute σ^2 is to gain use a shortcut formula.

PROPOSITION

$$V(X) = E(X^2) - [E(X)]^2$$

Example 4.12 (Example 4.10 continued)

For X = weekly gravel sales, we computed $E(X) = \frac{3}{8}$. Since

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx = \int_0^1 x^2 \cdot \frac{3}{2}(1 - x^2)\, dx$$
$$= \int_0^1 \frac{3}{2}(x^2 - x^4)\, dx = \frac{1}{5}$$
$$V(X) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = \frac{19}{320} = .059 \qquad \text{and} \qquad \sigma_X = .244$$

■

When $h(X) = aX + b$, the expected value and variance of $h(X)$ satisfy the same properties as in the discrete case: $E[h(X)] = a\mu + b$ and $V[h(X)] = a^2 \cdot \sigma^2$.

EXERCISES Section 4.2 (11–27)

11. The cdf of checkout duration X as described in Exercise 1 is

$$F(x) = \begin{cases} x < 0 \\ \frac{x^2}{4} \quad 0 \le x < 2 \\ 1 \quad 2 \le x \end{cases}$$

Use this to compute the following:

a. $P(X \le 1)$
b. $P(.5 \le X \le 1)$
c. $P(X > .5)$
d. The median checkout duration $\tilde{\mu}$ [solve $.5 = F(\tilde{\mu})$]
e. $F'(x)$ to obtain the density function $f(x)$

Works Cited

Devore, Jay L. Probability & Statistics for Engineering and the Sciences. 7th ed. CA: Thompson Brooks/Cole, 2008.