The Normal Distribution

The most important continuous probability distribution in the entire field of statistics is the **normal distribution**. Its graph, called the **normal curve**, is the bell-shaped curve of Figure 6.2, which approximately describes many phenomena that occur in nature, industry, and research. For example, physical measurements in areas such as meteorological experiments, rainfall studies, and measurements of manufactured parts are often more than adequately explained with a normal distribution. In addition, errors in scientific measurements are extremely well approximated by a normal distribution. In 1733, Abraham DeMoivre developed the mathematical equation of the normal curve. It provided a basis from which much of the theory of inductive statistics is founded. The normal distribution is often referred to as the **Gaussian distribution**, in honor of Karl Friedrich Gauss

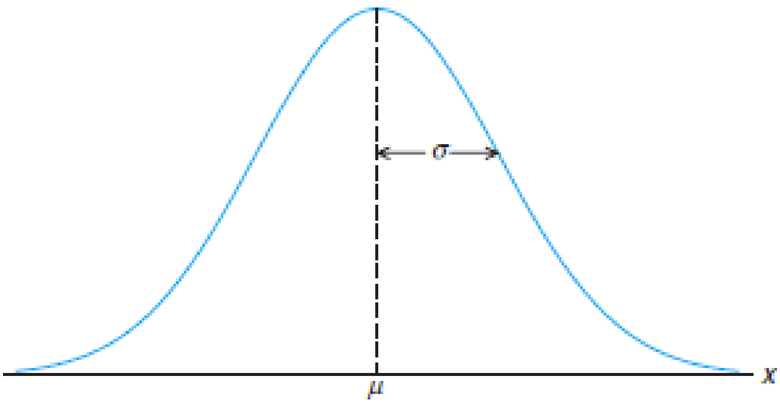

Figure 6.2: The normal curve.

(1777–1855), who also derived its equation from a study of errors in repeated measurements of the same quantity.

A continuous random variable X having the bell-shaped distribution of Figure 6.2 is called a **normal random variable**. The mathematical equation for the probability distribution of the normal variable depends on the two parameters μ and σ, its mean and standard deviation, respectively. Hence, we denote the values of the density of X by $n(x; \mu, \sigma)$.

Normal Distribution The density of the normal random variable X, with mean μ and variance σ^2, is

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

where $\pi = 3.14159\ldots$ and $e = 2.71828\ldots$.

Once μ and σ are specified, the normal curve is completely determined. For example, if $\mu = 50$ and $\sigma = 5$, then the ordinates $n(x; 50, 5)$ can be computed for various values of x and the curve drawn. In Figure 6.3, we have sketched two normal curves having the same standard deviation but different means. The two curves are identical in form but are centered at different positions along the horizontal axis.

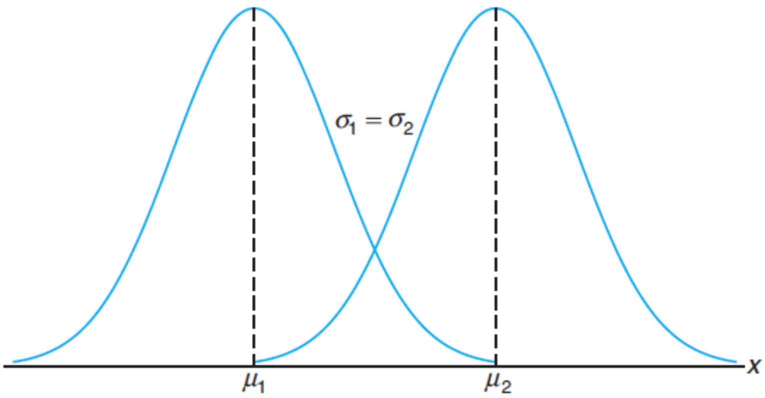

Figure 6.3: Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 = \sigma_2$.

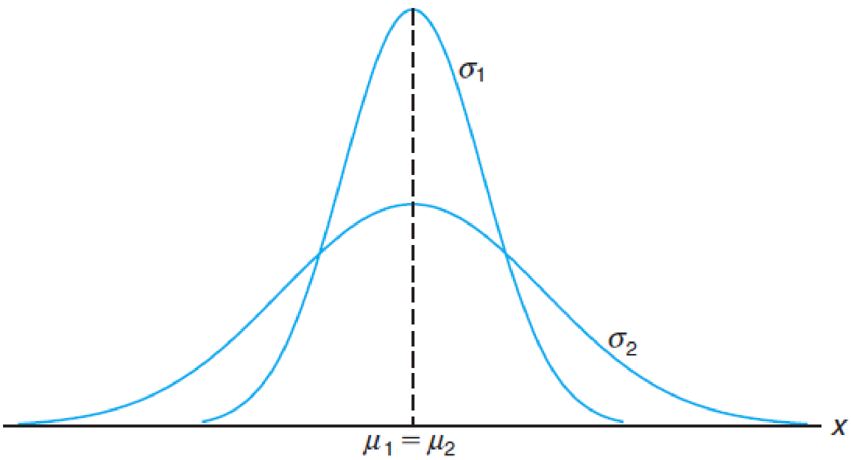

Figure 6.4: Normal curves with $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$.

In Figure 6.4, we have sketched two normal curves with the same mean but different standard deviations. This time we see that the two curves are centered at exactly the same position on the horizontal axis, but the curve with the larger standard deviation is lower and spreads out farther. Remember that the area under a probability curve must be equal to 1, and therefore the more variable the set of observations, the lower and wider the corresponding curve will be.

Figure 6.5 shows two normal curves having different means and different standard deviations. Clearly, they are centered at different positions on the horizontal axis and their shapes reflect the two different values of σ.

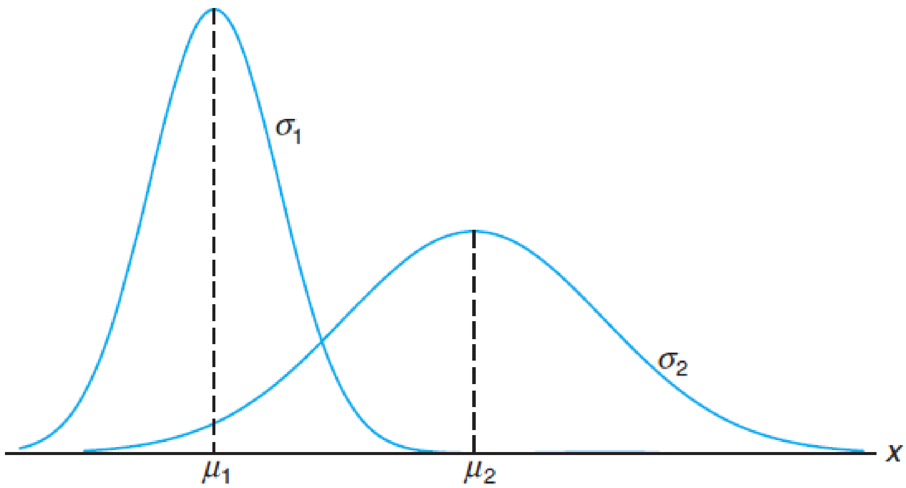

Figure 6.5: Normal curves with $\mu_1 < \mu_2$ and $\sigma_1 < \sigma_2$.

Based on inspection of Figures 6.2 through 6.5 and examination of the first and second derivatives of $n(x; \mu, \sigma)$, we list the following properties of the normal curve:

1. The mode, which is the point on the horizontal axis where the curve is a maximum, occurs at $x = \mu$.
2. The curve is symmetric about a vertical axis through the mean μ.
3. The curve has its points of inflection at $x = \mu \pm \sigma$; it is concave downward if $\mu - \sigma < X < \mu + \sigma$ and is concave upward otherwise.
4. The normal curve approaches the horizontal axis asymptotically as we proceed in either direction away from the mean.
5. The total area under the curve and above the horizontal axis is equal to 1.

Theorem 6.2:

The mean and variance of $n(x; \mu, \sigma)$ are μ and σ^2, respectively. Hence, the standard deviation is σ.

Area Under the normal curve:

The curve of any continuous probability distribution or density function is constructed so that the area under the curve bounded by the two ordinates $x = x_1$ and $x = x_2$ equals the probability that the random variable X assumes a value between $x = x_1$ and $x = x_2$. Thus, for the normal curve in Figure 6.6,

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma)\, dx = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

is represented by the area of the shaded region.

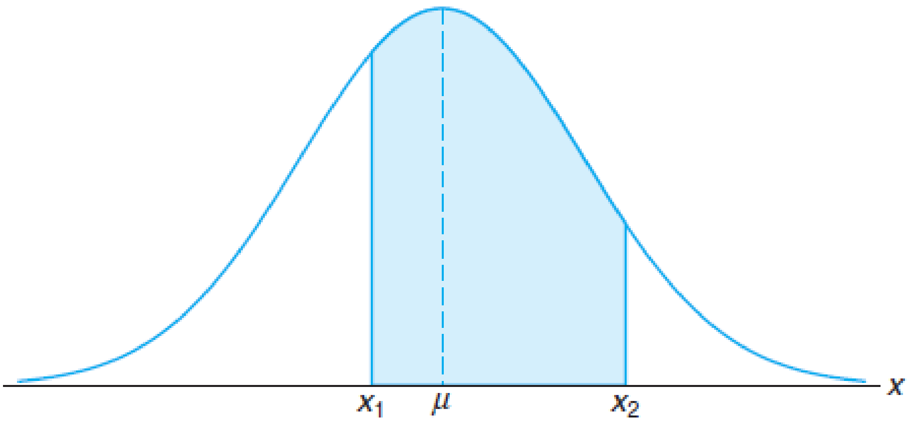

Figure 6.6: $P(x_1 < X < x_2)$ = area of the shaded region.

In Figures 6.3, 6.4, and 6.5 we saw how the normal curve is dependent on the mean and the standard deviation of the distribution under investigation. The area under the curve between any two ordinates must then also depend on the values μ and σ. This is evident in Figure 6.7, where we have shaded regions corresponding to $P(x_1 < X < x_2)$ for two curves with different means and variances. $P(x_1 < X < x_2)$, where X is the random variable describing distribution A, is indicated by the shaded area below the curve of A. If X is the random variable describing distribution B, then $P(x_1 < X < x_2)$ is given by the entire shaded region. Obviously, the two shaded regions are different in size; therefore, the probability associated with each distribution will be different for the two given values of X.

There are many types of statistical software that can be used in calculating areas under the normal curve. The difficulty encountered in solving integrals of normal density functions necessitates the tabulation of normal curve areas for quick reference. However, it would be a hopeless task to attempt to set up separate tables for every conceivable value of μ and σ. Fortunately, we are able to transform all the observations of any normal random variable X into a new set of observations

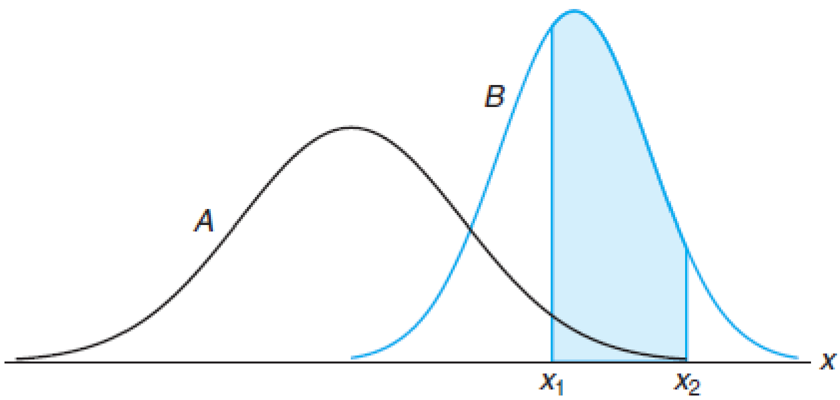

Figure 6.7: $P(x_1 < X < x_2)$ for different normal curves.

of a normal random variable Z with mean 0 and variance 1. This can be done by means of the transformation

$$Z = \frac{X - \mu}{\sigma}.$$

Whenever X assumes a value x, the corresponding value of Z is given by $z = (x - \mu)/\sigma$. Therefore, if X falls between the values $x = x_1$ and $x = x_2$, the random variable Z will fall between the corresponding values $z_1 = (x_1 - \mu)/\sigma$ and $z_2 = (x_2 - \mu)/\sigma$. Consequently, we may write

$$\begin{aligned} P(x_1 < X < x_2) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2} \, dz \\ &= \int_{z_1}^{z_2} n(z; 0, 1) \, dz = P(z_1 < Z < z_2), \end{aligned}$$

where Z is seen to be a normal random variable with mean 0 and variance 1.

Definition 6.1:

The distribution of a normal random variable with mean 0 and variance 1 is called a **standard normal distribution**.

The original and transformed distributions are illustrated in Figure 6.8. Since all the values of X falling between x_1 and x_2 have corresponding z values between z_1 and z_2, the area under the X-curve between the ordinates $x = x_1$ and $x = x_2$ in Figure 6.8 equals the area under the Z-curve between the transformed ordinates $z = z_1$ and $z = z_2$.

We have now reduced the required number of tables of normal-curve areas to one, that of the standard normal distribution. Table A.3 indicates the area under the standard normal curve corresponding to $P(Z < z)$ for values of z ranging from -3.49 to 3.49. To illustrate the use of this table, let us find the probability that Z is less than 1.74. First, we locate a value of z equal to 1.7 in the left column; then we move across the row to the column under 0.04, where we read 0.9591. Therefore, $P(Z < 1.74) = 0.9591$. To find a z value corresponding to a given probability, the process is reversed. For example, the z value leaving an area of 0.2148 under the curve to the left of z is seen to be -0.79.

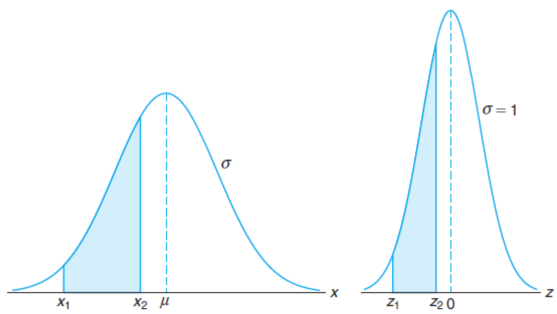

Figure 6.8: The original and transformed normal distributions.