# BASIC STATISTICS:

# A USER ORIENTED APPROACH

(*Manuscript*)

**Spyros MAKRIDAKIS**
**and**
**Robert L. WINKLER**

## CHAPTER 5

### 5.4 The Normal Distribution

A histogram representing the student ages used in Chapter 2 (see Table 2.5) is shown in Figure 5.15. In Figure 5.16, a bell-shaped curve known as the normal curve is superimposed on the histogram, and you can see that the histogram is very close to the curve. The curve represents the normal distribution, which is widely used in statistics. Not only ages, but many other variables, such as heights, weights, IQ scores, and exam grades, often have distributions that follow the normal curve quite closely. Moreover, an important result known as the Central Limit Theorem, which will be discussed in Chapter 7, provides additional justification for the application of the normal distribution in statistics.

A normal distribution can be specified by giving two values, the mean and the standard deviation of the distribution. The curve is bell-shaped, and the "bell" is centered at the mean. Notice that the curve is symmetric; the half to the right of the mean is a perfect mirror image of the half to the left of the mean. As the standard deviation changes, the "bell" becomes taller and thinner (for smaller standard deviations) or shorter and fatter (for larger standard deviations). Some examples are shown in Figure 5.17.

The formula for the normal curve is

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

where $\mu$ is the mean of the distribution and $\sigma$ is the standard deviation. As noted in the discussion of the Poisson distribution, the symbol e represents a constant, the base of the natural logarithm system, which is approximately 2.718. The symbol $\pi$ represents another constant, approximately 3.1416, which you may have encountered in geometry in formulas for the area and circumference of a circle.

Fortunately, although the formula for the normal curve is included here for completeness, you will not need to use this formula. Normal probabilities can be found from a table or a computer. Since the normal curve is continuous, probability is represented by area under the curve. The total area under the normal curve is one. Furthermore, for any normal distribution, the area under the curve between $\mu-\sigma$ and $\mu+\sigma$ is approximately 0.68 (Figure 5.18), the area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 0.95 (Figure 5.19), and the area between $\mu-3\sigma$ and $\mu+3\sigma$ is slightly greater than 0.997 (Figure 5.20). Thus, about 68 percent of the probability is within one standard deviation from the mean, 95 percent is within two standard deviations, and 99.7 percent is within three standard deviations. Even though the normal formula holds for any value of x, no matter how large or small, values of x more than three standard deviations from the mean are quite unlikely to occur.

To be able to find probabilities for normal distributions, first you need to concentrate on a special normal distribution, the standard normal distribution. This is a normal distribution with mean zero and standard deviation one. Because a standardized random variable, or a standard score, is usually denoted by z, z is used here to represent a variable with a standard normal distribution. Table can be used to find probabilities for z.

In Table 3, cumulative probabilities such as $P(z > 1.52)$ are given. To find $P(z > 1.52)$, look up $z = 1.52$ and read the probability, $P(z > 1.52) = 0.0643$. This probability is shown in Figure 5.21.

For negative values of z, the symmetry of the normal curve can be used. For example, the area to the left of -1.52 is the same as the area to the right of +1.52, as shown in Figure 5.22. But the area to the left of -1.52 is found by taking $P(z > -1.52) = 0.9357$ and subtracting it from one:

$$P(z < -1.52) = 1 - P(z > -1.52) = 1 - 0.9357 = 0.0643$$

Table 3 gives cumulative probabilities for both positive and negative values of z.

Often the probability that z is between two values is of interest. How can $P(0.43 < z < 1.75)$ be found? The area to the right of 1.75 is 0.0401, from Table 3, and the area to the right of 0.43 is 0.3336. Thus, the area between 0.43 and 1.75 must be 0.3336 - 0.0401 = 0.2935.

For a final example involving standard normal probabilities, let us find the shaded area in Figure 5.23, which represent $P(-1.5 < z < 0.5)$. From the normal table, the area to the right of 0.5 is 0.3085. The area to the right of -1.5 is 0.9332.

Subtracting the two values gives

$$P(-1.5 < z < 0.5) = 0.9332 - 0.3085 = 0.6247$$

When you are finding a normal probability, it is a good idea to draw a rough sketch to keep track of the area you want. Such a sketch reduces the chance of making simple mistakes such as finding the area to the left of a value when you really want the area to the right. You will be using the normal distribution very often in later chapters, so it is a good idea to become familiar with computing normal probabilities. Even though a computer can do this work for you, you should learn how to read a table of normal probabilities and how to find specific areas under the normal curve.

Probabilities for any normal distribution can be related to standard normal probabilities, since z can be related to x by subtracting μ from x and then dividing by σ:

$$z = \frac{x - \mu}{\sigma}.$$

For example, suppose that scores on a college entrance examination are normally distributed with mean 600 and standard deviation 50. What is the probability of a person scoring less than 640? This is $P(x < 640)$, but $x = 640$ corresponds to a standard score of

$$z = \frac{x - \mu}{\sigma} = \frac{640 - 600}{50} = 0.80.$$

Since an x less than 640 corresponds to a z less than 0.80,

$$P(x < 640) = P(z < 0.80) = 1 - P(x > 0.80) = 1 - 0.2119 = 0.7881$$

from Table 3. This probability is shown in Figure 5.24. Table 3 gives probabilities only for z, but these can be used to find probabilities for x from any normal distribution.

What proportion of students taking the college entrance examination receive scores between 580 and 680? Converting from x to z yields

$$P(580 < x < 680) = P\left(\frac{580 - \mu}{\sigma} < \frac{x - \mu}{\sigma} < \frac{680 - \mu}{\sigma}\right)$$

$$= P\left(\frac{580 - 600}{50} < z < \frac{680 - 600}{50}\right) = P(-0.40 < z < 1.60).$$

Next, from Table 3,

$$P(z > -0.40) = 0.6554,$$
$$P(z > 1.60 = 0.0548,$$

and $$P(z > -0.40) - P(z > 1.60) = 0.6554 - 0.0548 = 0.6006.$$

Therefore,

$$P(-0.40 < z < 1.60) = 0.6006.$$

Table 3 can also be used to find percentiles of a normal distribution. For instance, what is the 90th percentile of the distribution of scores on the college entrance examination? This is a value of x which has an area of 0.90 to the left and 0.10 to the right. The cumulative probability at x is 0.10. If you look at the cumulative probabilities in Table 3, you will find that the closest values to 0.10 are 0.1003 (for z = 1.28) and 0.0985 (for z = 1.29). Since 0.1003 is closer to 0.10 than 0.0985, we will take 1.28 as the 90th percentile of z. By manipulating

$$z = \frac{x - \mu}{\sigma}$$

algebraically, we get x as a function of z:

$$x = \mu + z\sigma.$$

Then, if 1.28 is the 90th percentile of z, the 90th percentile of x is

$$\mu + 1.28\sigma = 600 + 1.28(50) = 664.$$

This means that 90 percent of those taking the examination score less than 664. Any score above 664 is in the top 10 percent of the distribution of scores, as shown in Figure 5.25.

In Section 5.2 the Poisson approximation to the binomial distribution was discussed. The normal distribution can also be used to approximate the binomial distribution. The Poisson approximation is good when the number of trials n is large and the binomial probability p is small. The normal approximation is good when n is large and p is not too small or too large. One rough rule of thumb is to use the normal approximation when both $np \geq 5$ and $n(1-p) \geq 5$ are satisfied.

From Section 5.1, the mean and standard deviation of a binomial distribution are

$$\mu = np$$

and

$$\sigma = \sqrt{npq}.$$

For a normal approximation, we simply take a normal distribution with the same mean, np, and the same standard deviation, $\sqrt{npq}$. The z score corresponding to a given value of x is therefore

$$z = \frac{x - \mu}{\sigma} = \frac{x - np}{\sqrt{npq}}.$$

**Figure 5.15 : Histogram of Student Ages (See Table 2.1 for Data).**

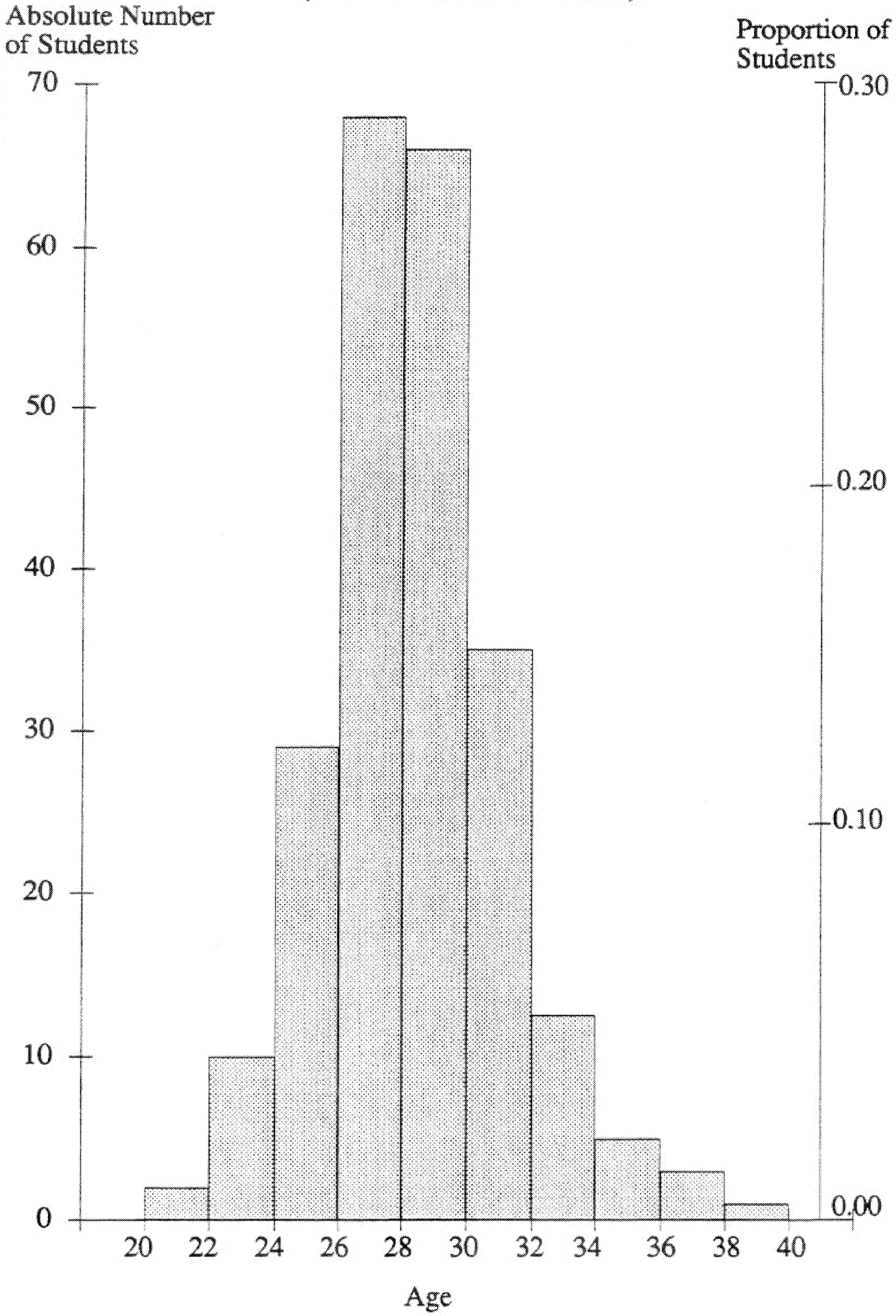

**Figure 5.16 : Histogram of Student Ages with Normal Curve.**

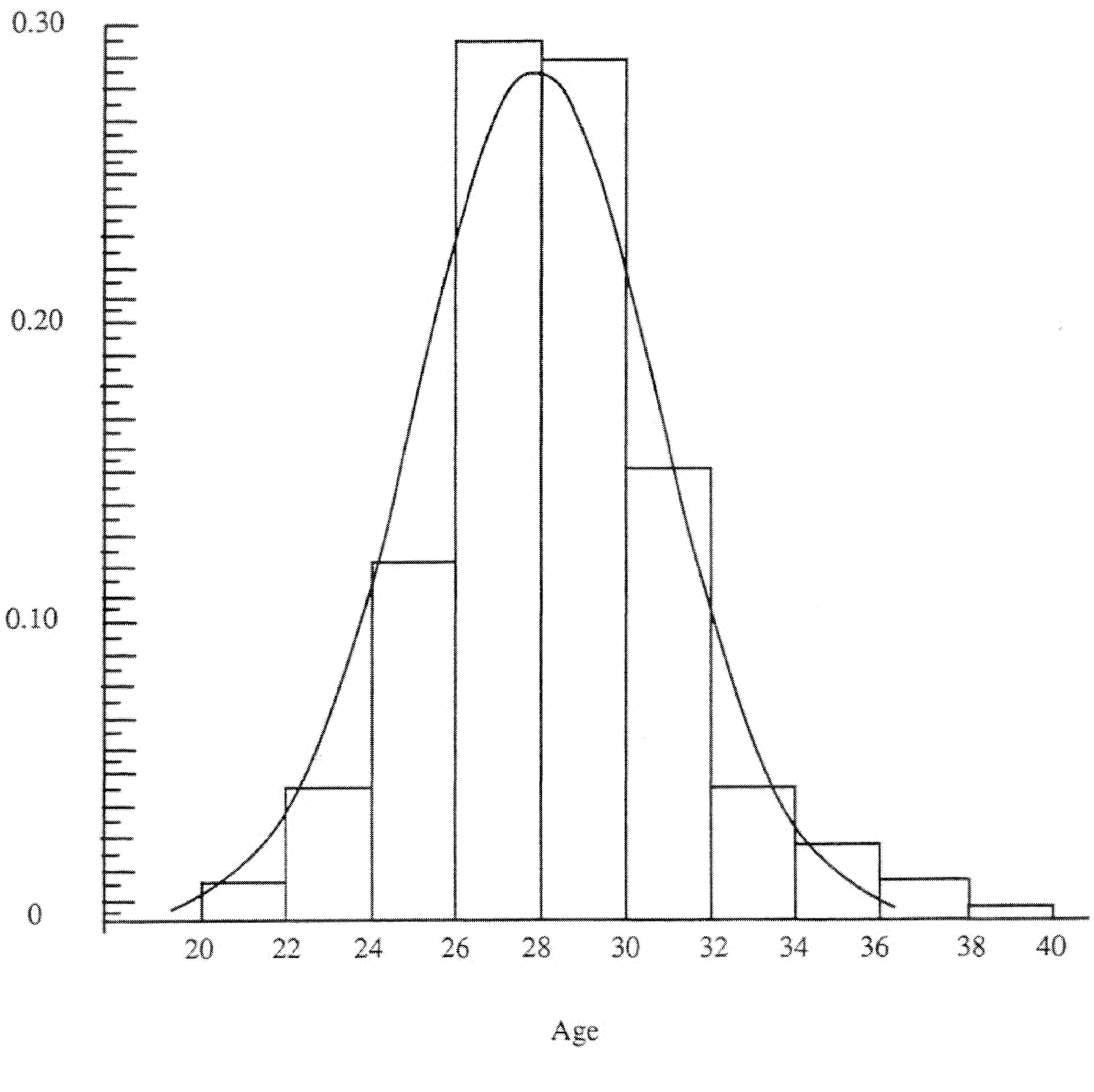

**Figure 5.17 : Some Normal Curves.**

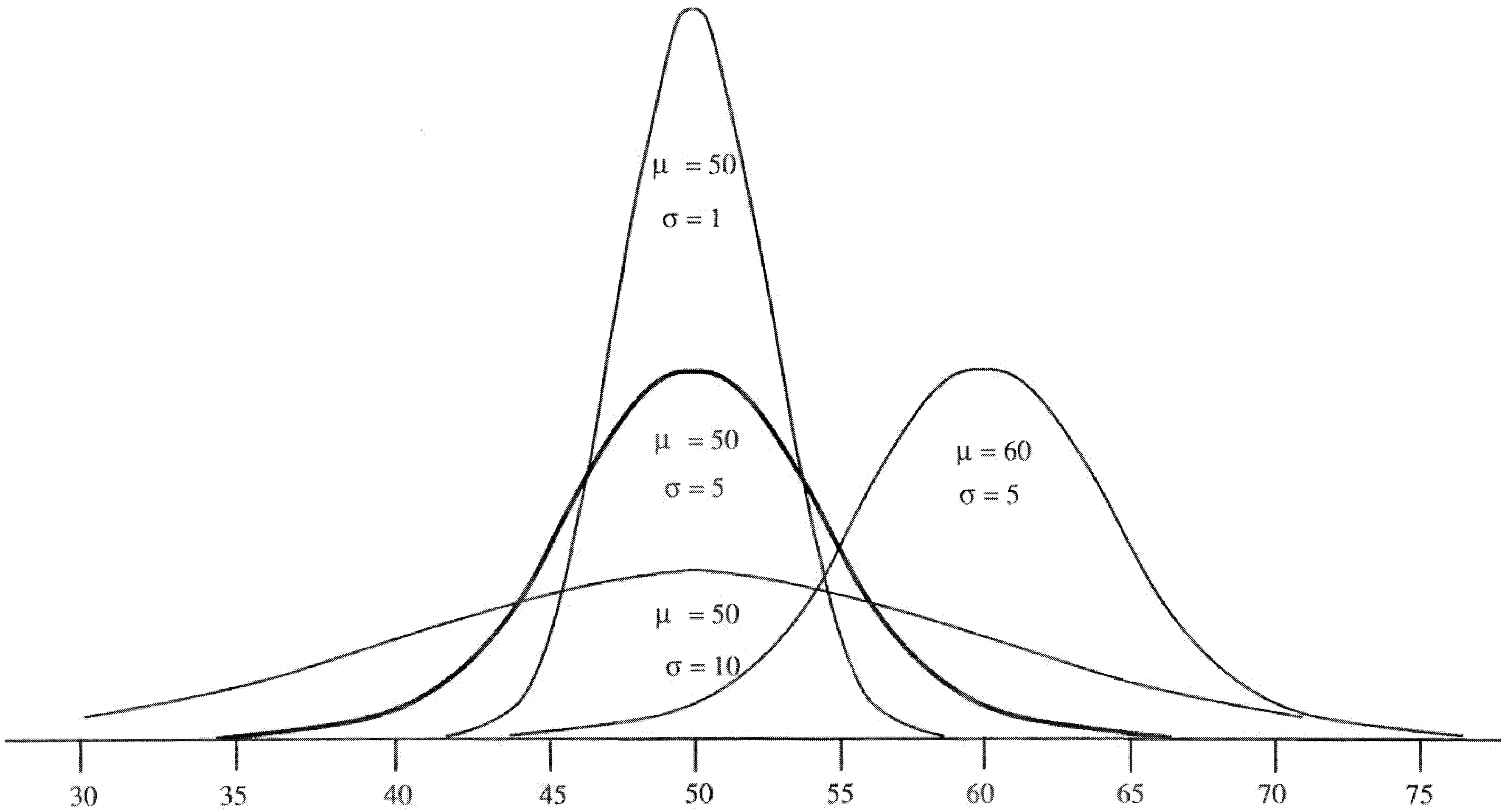

**Figure 5.18 : The Area within One Standard Deviation of the Mean in a Normal Curve.**

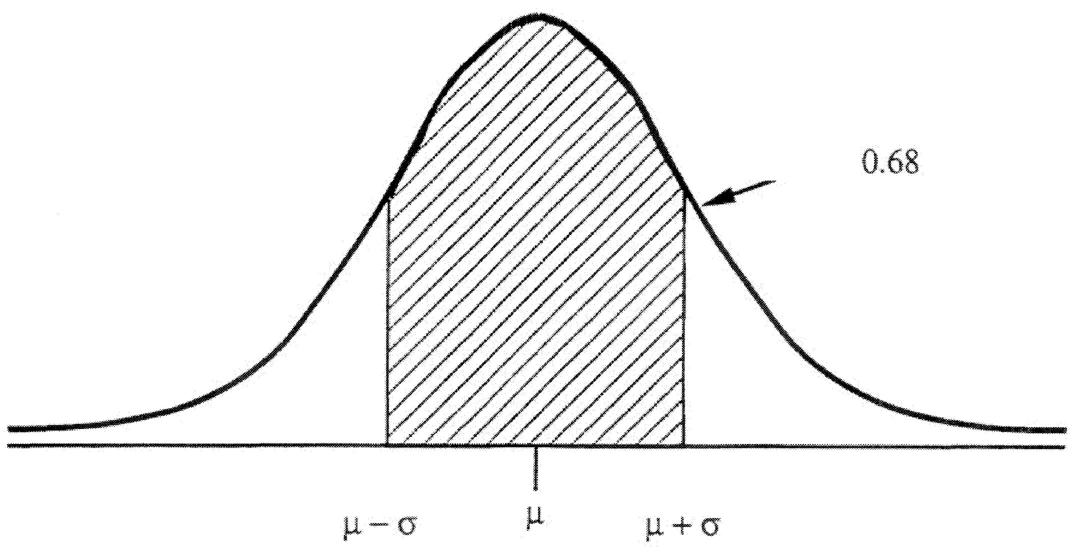

**Figure 5.19 : The Area within Two Standard Deviations of the Mean in a Normal Curve.**

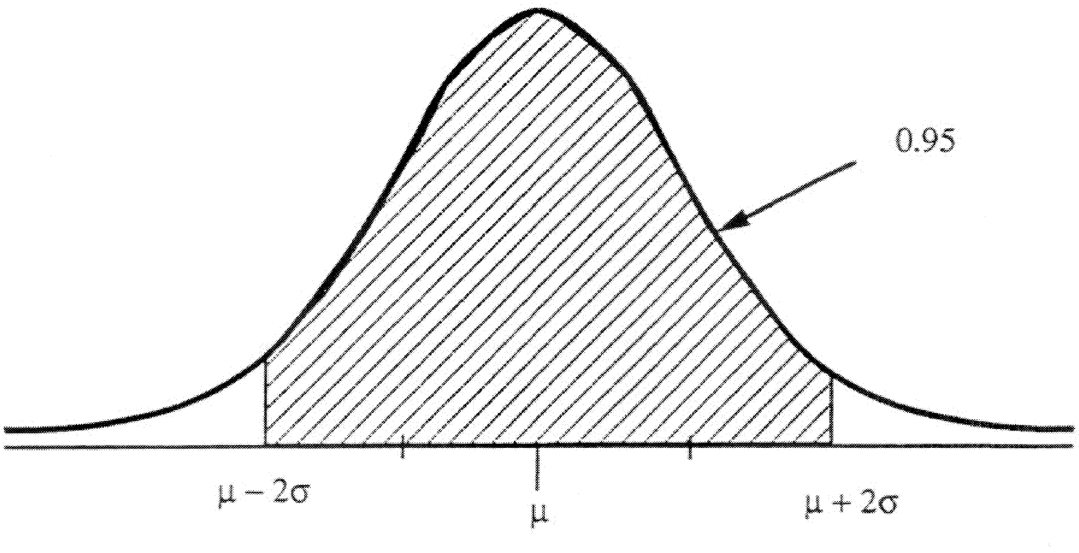

**Figure 5.20 : The Area within Three Standard Deviations of the Mean in a Normal Curve.**

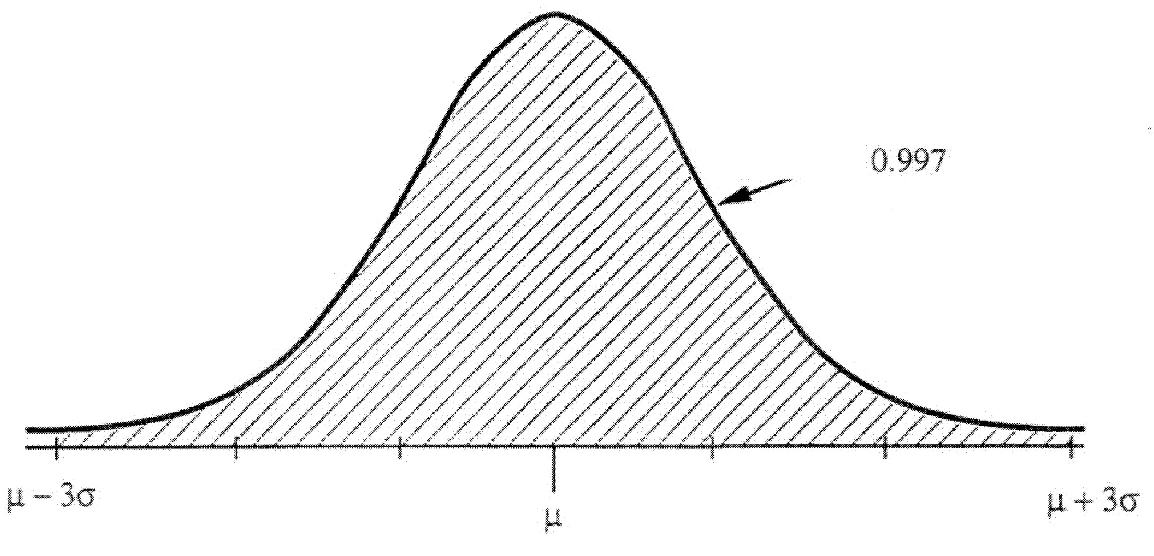

**Figure 5.21 : The Area to the Right of z = 1.52 in a Standard Normal Distribution.**

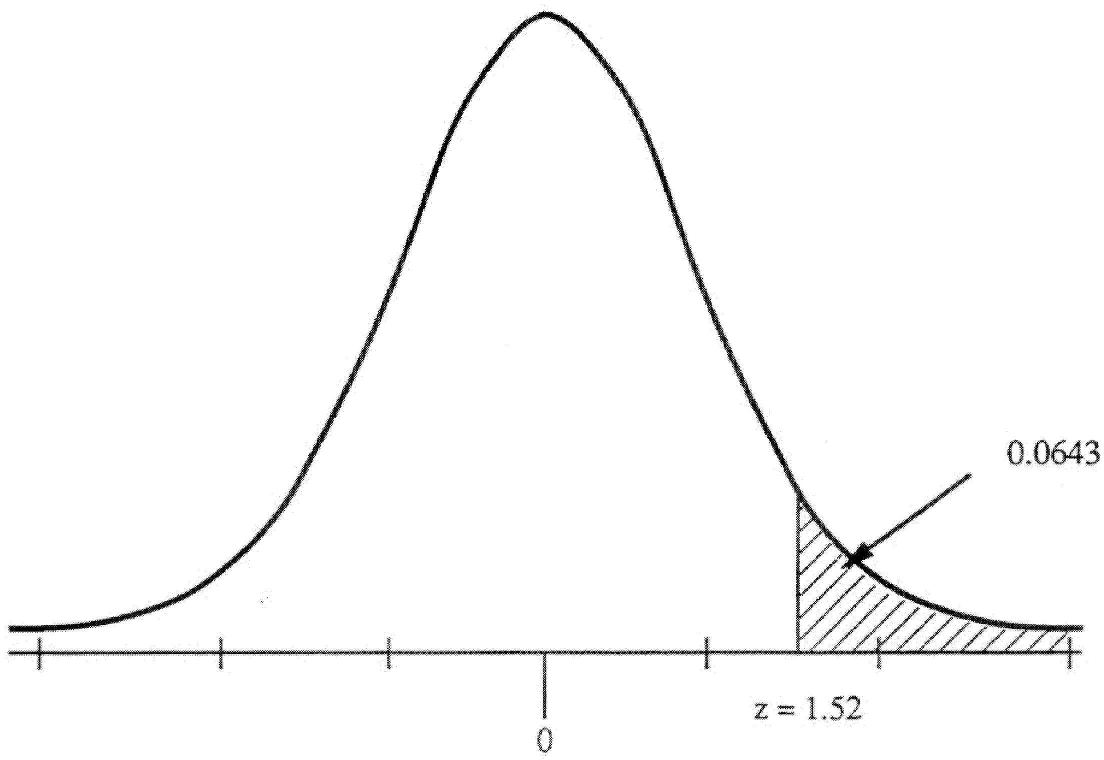

**Figure 5.22 : Equal Areas in the Tails of a Standard Normal Distribution.**

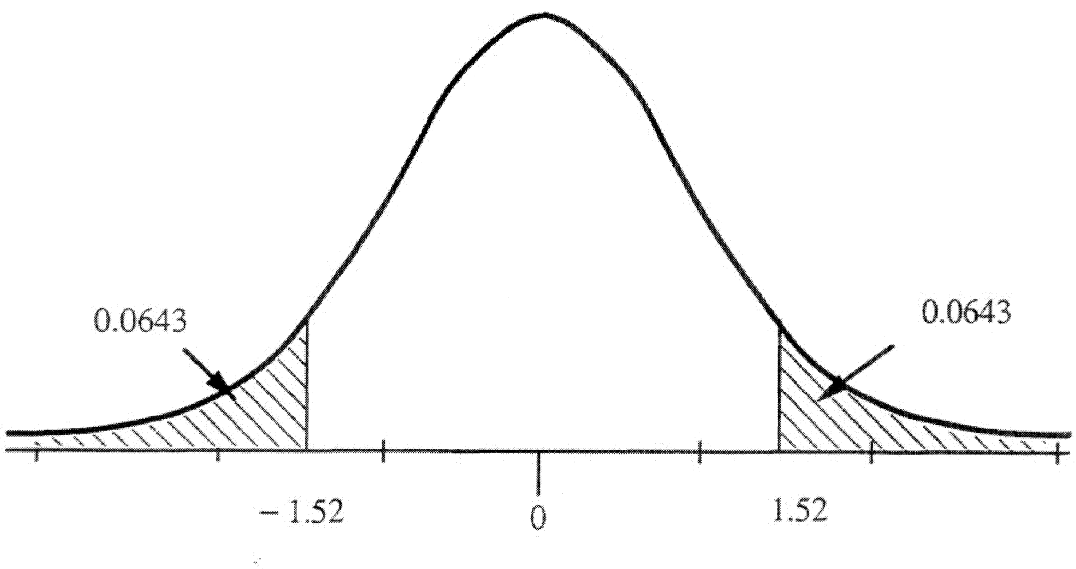

**Figure 5.23 : P ( - 1.5 < z < 0.5 ) .**

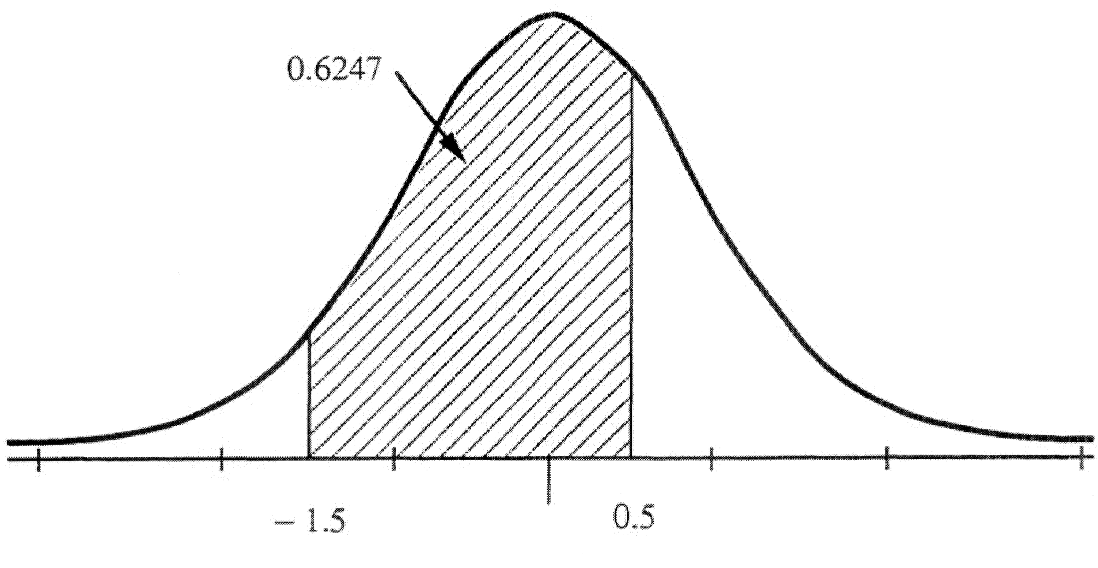

**Figure 5.24 : The Probability of a Score less than 640 when $\mu = 600$ and $\sigma = 50$.**

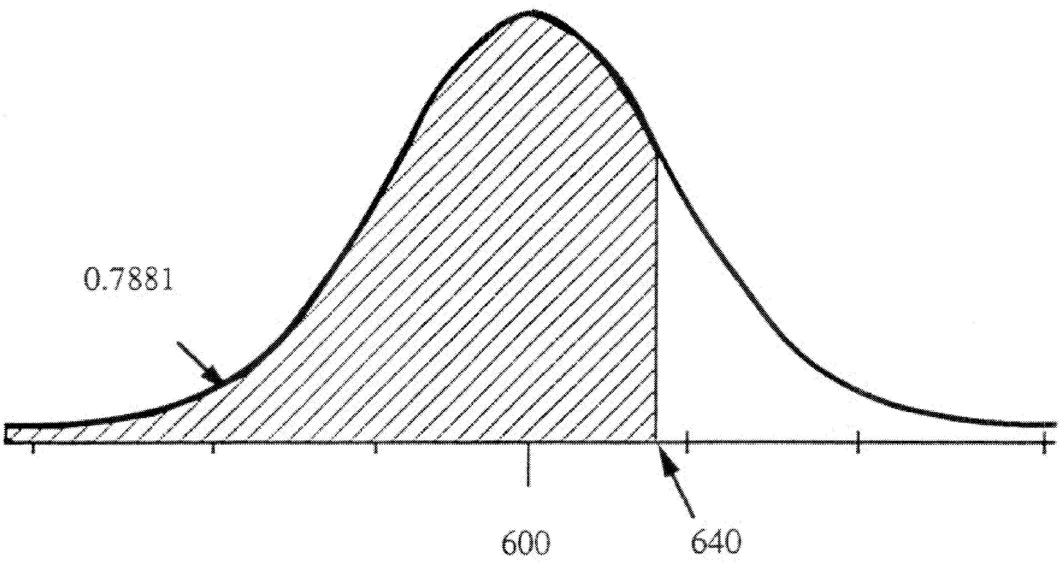

**Figure 5.25 : The 90th Percentile of the Normal Distribution with $\mu = 600$ and $\sigma = 50$.**