

Unit 15

Density Curves and z -Scores

Objectives:

- To understand how a density curve describes the distribution of a population
- To become acquainted with the properties of a normal density curve
- To compute and interpret z -scores and raw scores
- To recognize the important roll that the normal distribution plays with regard to the sampling distribution of \bar{x} with simple random sampling

The histogram and the frequency polygon are two graphical displays that we have encountered on several previous occasions. For instance, Figure 3-4 is a histogram displaying the distribution of "Number of Children" from the SURVEY DATA, displayed as Data Set 1-1 at the end of Unit 1, and Figures 3-3 and 3-6 are respectively a histogram and a frequency polygon displaying the distribution of "Yearly Income" from the SURVEY DATA. Each of these graphical displays is designed to display the distribution of a variable for a sample. Recall that using graphical displays to describe the characteristics of a sample is part of what we call descriptive statistics. However, our primary focus presently is on using a sample to draw conclusions about a population, which is what we call inferential statistics. In order to develop our study of inferential statistics, we now want to consider how to display the distribution of a variable for a population.

We have stated previously that, in most practical situations, we think of populations as containing a very large, possibly even infinite, number of items. For some variables, we are able to make an ordered list of the possible, observable values of a variable. For example, if we are studying the variable "Number of Children" in a very large population, an ordered list of the possible, observable values is the non-negative integers 0, 1, 2, 3, ..., etc.; we call such a list of possible values *discrete*. On the other hand, there are many variables for which such an ordered list is not possible. For example, if we are studying the variable "Height," non-integer values are possible. Someone may have a height of 68 inches, or a height of 69 inches, or a height of any number of inches between 68 and 69 (such as 68.34 inches). We are not able to make an ordered list of the possible, observable values, because no matter what list is constructed, we can always find a possible value in between any two values on the list. When we can always find a possible value in between any two values of a variable, we say that the values lie on a *continuum*, and we call the possible values *continuous*. The variable "Number of Children" would not be treated as continuous, because it is possible for someone to have 3 children, and it is possible for someone to have 4 children, but it is not possible for someone to have a number of children in between 3 and 4.

In very many practical situations, we can treat a variable which has a discrete list of possible values as if its possible values are continuous. For example, when are studying the variable "Yearly Income," one might say that the possible values are discrete, since it is possible for someone to have an income of \$36,345.67, and it is possible for someone to have an income of \$36,345.68, but it is not possible for someone to have an income in between \$36,345.67 and \$36,345.68. While strictly speaking, this is true, the list of possible incomes is so large, that it is not at all unreasonable to think of the possible incomes as being on a continuum. As a general rule, when the number of possible values for a variable is large, we can treat the variable as if it is continuous. In fact, many statistical procedures designed for continuous-type variables work well for discrete-type variables even if the number of possibilities is not really very large.

For this reason, our immediate focus will be on populations for which the possible values of the variable being studied are treated as continuous. With this in mind let us return to the variable "Yearly Income." If we were going to construct a histogram or frequency polygon to represent the distribution for a very large population, we would be able to define many interval classes of small length. Figures 3-3 and 3-6 are based on six classes each representing an interval of length 10 thousand dollars to display data consisting of 30 incomes. For a much larger number of incomes, we could use 12 classes each representing a length of 5 thousand dollars, or we could use 24 classes each representing a length of 2.5 thousand dollars, or we could use 48 classes each representing a length of 1.25 thousand dollars, etc.

Figure 15-1a
Uniform Density Curve



Figure 15-1b
Normal (Bell-Shaped) Density Curve

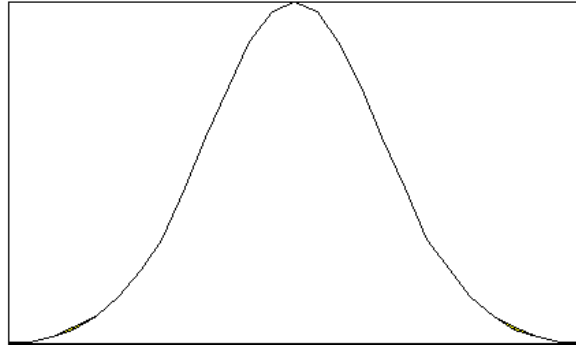


Figure 15-1c
Positively Skewed Density Curve

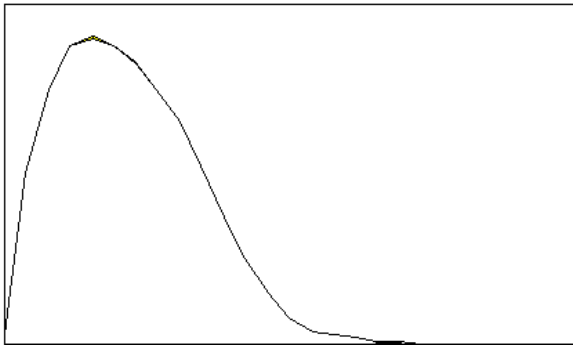


Figure 15-1d
Negatively Skewed Density Curve

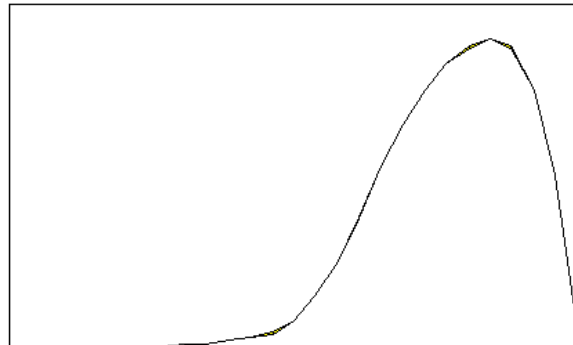


Figure 15-1e
Multimodal Density Curve

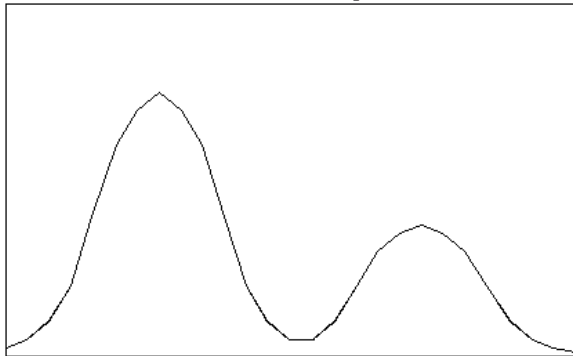
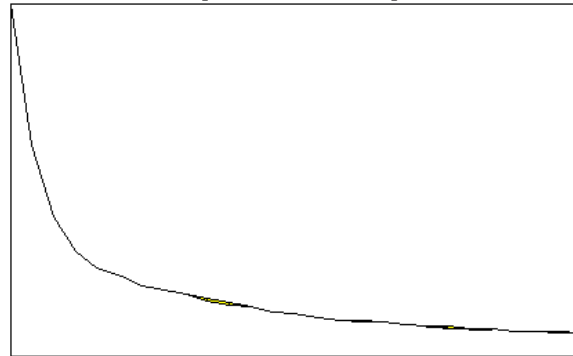


Figure 15-1f
Positively Skewed Density Curve



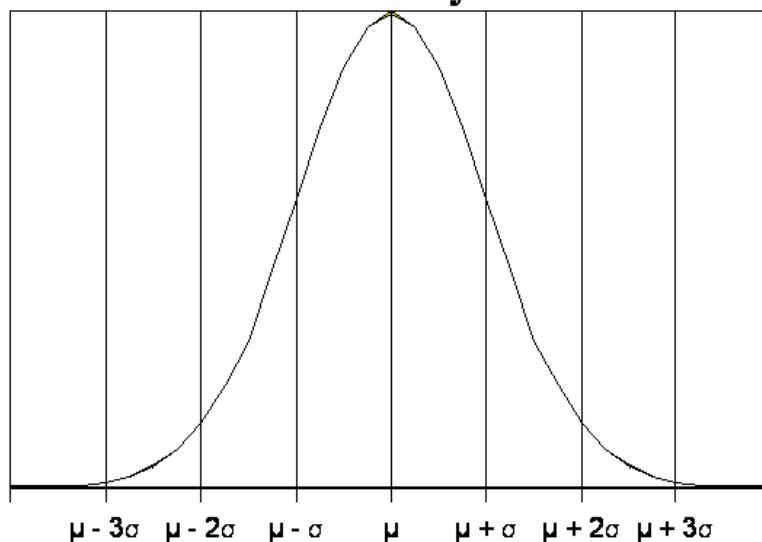
Increasing the number of intervals and decreasing the interval length will have a smoothing effect on the histogram or frequency polygon. In general, the distribution for a very large population can be represented as a smooth curve; such a curve is called a *density curve*. The proportion of area under a density curve between any two values on the horizontal axis represents the relative frequency of items which fall between the two given values; we may also interpret this relative frequency as the probability that one randomly selected item falls between the two given values. Figures 15-1a to 15-1f display several density curves, each illustrating characteristics similar to the corresponding histogram in Figures 3-5a to 3-5f.

Figure 15-1b and Figure 3-5b each display a symmetric, bell-shaped distribution. Notice that Figure 15-1b has been named the *normal density curve*, also called the *normal distribution*. This symmetric, bell-shaped distribution is given this special name because of the important role it plays in statistics with regard to sampling distributions. We have recently seen how the sampling distribution of \bar{x} with simple random sampling becomes more bell-shaped as the sample size n increases; stated more precisely, we can say that the sampling distribution becomes more like a normal distribution as the sample size increases. Consequently, we shall need to study the normal distribution in detail. (The name *normal distribution* might be somewhat misleading, since it can give the impression that other distributions are somehow abnormal, which is not the case. The name *normal distribution* refers to a mathematical operation called normalizing.)

A population having a normal distribution will be most conveniently described in terms of its mean and standard deviation. Recall that we use μ to represent the mean of a population and \bar{x} to represent the mean of a sample. We shall now need to distinguish between the standard deviation of a population and the standard deviation of a sample; we shall use σ (the lower case Greek letter sigma) to represent the standard deviation of population, and we shall use s to refer exclusively to the standard deviation of a sample. In other words, μ and σ are parameters, and \bar{x} and s are statistics.

A normal distribution is symmetric and can be described entirely in terms of the number of standard deviations values are way from the mean. Figure 15-2 displays a more descriptive density curve than Figure 15-1b. The area under the curve has been marked in terms of one, two, and three standard deviations from the mean. One of the properties of a normal distribution, which can be seen in Figure 15-2, is that well over half of the items in a normally distributed population are within one standard deviation of the mean; more specifically, 68.3% of the area under a normal density curve is between $\mu - \sigma$ and $\mu + \sigma$. You should also notice from Figure 15-2 that most the items in a normally distributed population are within two standard deviations of the mean; more specifically, 95.4% of the area under a

Figure 15-2
Normal Density Curve



normal density curve is between $\mu - 2\sigma$ and $\mu + 2\sigma$. Finally, we notice from Figure 15-2 that practically all of the items in a normally distributed population are within three standard deviations of the mean. Even though a normal density curve extends infinitely in both directions without ever touching the horizontal axis, the amount of area under the curve outside three standard deviations is negligible; more specifically, 99.7% of the area under a normal density curve is between $\mu - 3\sigma$ and $\mu + 3\sigma$, which leaves only slightly more than one quarter of one percent of the area outside three standard deviations.

Figures 15-3a and 15-3b display normal distributions each representing the distribution of yearly salaries of employees from one of two fictitious corporations named Alco and Balsa. Even though both distributions are normal, they have very different means and standard deviations. For the Alco corporation, $\mu = \$20,000$ and $\sigma = \$2,000$; for the Balsa corporation, $\mu = \$25,000$ and $\sigma = \$4,000$. The area under each curve has been marked in terms of one, two, and three standard deviations from the mean. Also, the two normal density curves have been displayed underneath one another and on the same scale so that the differences between the two distributions will be highlighted.

The fact that the normal density curve for Balsa is to the right of the normal density curve for Alco is a reflection of the fact that the mean salary is greater at Balsa than at Alco. The fact that the normal density curve for Balsa is more spread out than the normal density curve for Alco is a reflection of the fact that standard deviation of salaries is greater at Balsa than at Alco (i.e., there is more variation in salaries at Balsa). The mean

μ , which is a measure of the center of a distribution, determines where the normal density curve will be centered; the standard deviation σ , which is a measure of dispersion, determines how spread out the normal density curve will be.

Suppose George is an employee of Alco and earns \$22,000; also, suppose Sandy is an employee of Balsa and earns \$27,000. Even though Sandy earns more money than George, we shall see that George has a higher salary than Sandy relative to the corporations for which they work. Observe that each of these two employees has a salary which is \$2,000 higher than the mean salary of his/her respective corporation. Locate George's salary of \$22,000 on the horizontal axis of Figure 15-3a; then locate Sandy's salary of \$27,000 on the horizontal axis of Figure 15-3b. You should realize that the area above George's \$22,000 salary under the normal curve for Alco is less than the area above Sandy's \$27,000 salary under the normal curve for Balsa. This is an indication that the percentage of employees at Balsa earning more than Sandy is greater than the percentage of employees at Alco earning more than George, which suggests that George has a higher salary than Sandy relative to the corporations for which they work, even though Sandy earns more money than George. The reason why this happens is related to the fact that salaries at Balsa are more dispersed than salaries at Alco.

Once again locating George's salary of \$22,000 on the horizontal axis of Figure 15-3a, you should see that George's salary is exactly one whole standard deviation above the mean; however, once again locating Sandy's salary of \$27,000 on the horizontal axis of Figure 15-3b, you should see that Sandy's salary is exactly 1/2 of one standard deviation above the mean. Earlier, we stated that 68.3% of the items in a normally distributed population are between $\mu - \sigma$ and $\mu + \sigma$. This implies that 68.3% of the salaries at the Alco corporation are between \$18,000 and \$22,000, and that 68.3% of the salaries at the Balsa corporation are between \$21,000 and \$29,000. Consequently, $100\% - 68.3\% = 31.7\%$ of the salaries at the Alco corporation are outside the range from \$18,000 and \$22,000, and 31.7% of the salaries at the Balsa corporation are outside the range from \$21,000 and \$29,000. Since a normal distribution is symmetric, 15.85% (half of 31.7%) of the salaries at the Alco corporation are above \$22,000, and 15.85% of the salaries at the Balsa corporation are above \$29,000. We see then that $100\% - 15.85\% = 84.15\%$ of the salaries at the Alco corporation are below George's \$22,000 salary, whereas less than 84.15% of the salaries at the Balsa corporation are below Sandy's \$27,000 salary.

In general, if two values come from two different populations (or samples), we can compare the two values relative to the populations (samples) from which they come by considering the number of standard deviations each value is away from its respective population (sample) mean. The number of standard deviations a value is away from its population mean is called a *z-score*, or *standard score*. If x represents the value of a

Figure 15-3a

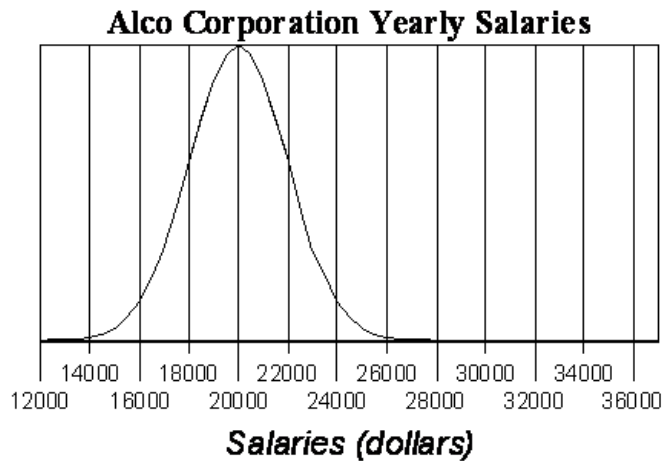
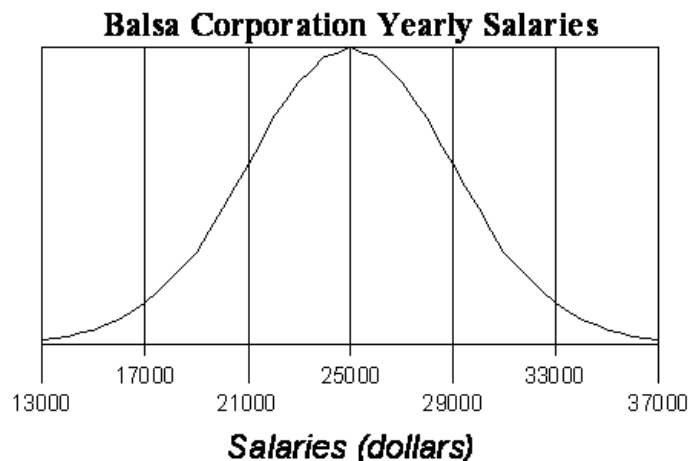


Figure 15-3b



variable corresponding to a population of values with mean μ and standard deviation σ , then the z -score, or standard score, of x is

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

The z -score of x is the number of standard deviations (σ s) x is away from the mean (μ); the z -score is negative when x is less than the mean and is positive when x is larger than the mean. Although z -scores can be used with any distribution, they are primarily used with regard to normal distributions, since as we shall soon see, they are important in describing normal distributions.

The z -score of George's \$22,000 salary at Alco is $(22,000 - 20,000) / 2000 = +1.00$, which simply confirms that George's salary is exactly one standard deviation above the mean. The z -score of Sandy's \$27,000 salary at Balsa is $(27,000 - 25,000) / 4000 = +0.50$, which simply confirms that Sandy's salary is exactly one half of one standard deviation above the mean.

As another illustration, let us compare Carolyn's \$17,000 salary at Alco with Will's \$19,000 salary at Balsa. The z -score of Carolyn's \$17,000 salary at Alco is $(17,000 - 20,000) / 2000 = -1.50$, which indicates that Carolyn's salary is 1.50 standard deviations below the mean. The z -score of Will's \$19,000 salary at Balsa is $(19,000 - 25,000) / 4000 = -1.50$, which indicates that Will's salary is 1.50 standard deviations below the mean. Consequently, the percentage of employees at Alco earning less than Carolyn is the same as the percentage of employees at Balsa earning less than Will; this suggests that Carolyn and Will have the same salary relative to the corporations for which they work, even though Will earns more money than Carolyn.

Earlier, we stated that 68.3% of the items in a normally distributed population are between $\mu - \sigma$ and $\mu + \sigma$, and that 95.4% of the items in a normally distributed population are between $\mu - 2\sigma$ and $\mu + 2\sigma$. This implies that 68.3% of the salaries at Alco are between \$18,000 and \$22,000, and that 95.4% of the salaries at Alco are between \$16,000 and \$24,000. Consequently, $100\% - 68.3\% = 31.7\%$ of the salaries at Alco are outside the range from \$18,000 and \$22,000, and $100\% - 95.4\% = 4.6\%$ of the salaries at Alco are outside the range from \$16,000 and \$24,000. Since a normal distribution is symmetric, 15.85% (i.e., half of 31.7%) of the salaries at Alco are below \$18,000, and 2.3% (i.e., half of 4.6%) of the salaries at Alco are below \$16,000. We see then that somewhere between 2.3% and 15.85% of the salaries at Alco are below Carolyn's \$17,000 salary.

Will's \$19,000 salary at Balsa is between one and two standard deviations below the mean, just as Carolyn's is at Alco. Therefore, we can say that somewhere between 2.3% and 15.85% of the salaries at Balsa are below Will's \$19,000 salary.

We use $z = (x - \mu) / \sigma$ to convert the value x into a z -score; the value x can be called a *raw score*. There may be times when we want to convert a z -score back into a raw score. By solving the equation $z = (x - \mu) / \sigma$ for x , we find that

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

which we can use to convert a z -score into a raw score.

Suppose, for instance, that Ed's salary at the Balsa corporation is \$20,000, and we want to find what the comparable salary would be at the Alco corporation. First, we find that the z -score of Ed's \$20,000 salary at Balsa is $(20,000 - 25,000) / 4000 = -1.25$. In order to find a comparable salary at Alco, we need to find the salary whose z -score is -1.25 . This is easily done by converting -1.25 back into a raw score. The salary at Alco which is comparable to Ed's \$20,000 salary at Balsa is $\$20,000 + (-1.25)(\$2000) = \$17,500$.

We have described how the areas under a normal density curve are distributed only with regard to one, two, and three standard deviations away from the mean. Very shortly, we shall see that a more precise description of a normal distribution is available. When working with normal distributions, we focus on ranges of values of a variable rather than with a specific value of a variable. For instance, with the distribution of salaries at the Alco corporation, displayed in Figure 15-3a, we might focus on the percentage of salaries between \$18,000 and \$23,000, or the percentage of salaries less than \$21,000, but there is no point to considering the percentage of salaries exactly equal to \$19,000. The reason we do not focus on specific values is because the normal distribution is used to describe variables with values on a continuum. With values on a continuum, our interest is in intervals rather than in specific values.

We have already noted how the sampling distribution of \bar{x} with simple random sampling becomes more bell-shaped, like a normal distribution, with increasing sample size n . After discussing normal distributions in more detail, we shall see specifically how a normal distribution can be used to describe the sampling distribution of \bar{x} when n is sufficiently large.

Self-Test Problem 15-1. The time that it takes third graders to solve a puzzle is normally distributed with a mean of 130 seconds and a standard deviation of 8 seconds. The time that it takes fifth graders to solve the puzzle is normally distributed with a mean of 110 seconds and a standard deviation of 4 seconds.

- About what percentage of fifth graders can solve the puzzle in between 106 and 114 seconds?
- In what range can we be sure that the time to solve the puzzle will fall for at least 99% of all third graders will be?
- About what percentage of fifth graders can solve the puzzle in less than 118 seconds?
- About what percentage of third graders can solve the puzzle in more than 138 seconds?
- What can you say about the percentage of fifth graders who can solve the puzzle in more than 120 seconds?
- What can you say about the percentage of third graders who can solve the puzzle in less than 100 seconds?
- How does a fifth grader who can solve the puzzle in 115 seconds compare with a third grader who can solve the puzzle in 136 seconds relative to the grades which each is in?
- How does a fifth grader who can solve the puzzle in 107 seconds compare with a third grader who can solve the puzzle in 107 seconds relative to the grades which each is in?
- How much time for a third grader to solve the puzzle would be comparable to a time of 116 seconds for a fifth grader to solve the puzzle?
- How much time for a fifth grader to solve the puzzle would be comparable to a time of 118 seconds for a third grader to solve the puzzle?

Self-Test Problem 15-2. Suppose we know that the mean yearly income per household in a particular state is \$40,000 and that the standard deviation is \$12,000.

- What might we conclude about the distribution of incomes in the state if we found that almost 5% of the households in the state have a yearly income greater than \$76,000?
- Is it reasonable to assume that practically all yearly household incomes are between \$40,000 and \$76,000? Why or why not?

Answers to Self-Test Problems

- 15-1** (a) 68.3% (b) between 106 and 154 seconds (c) 97.7% (d) 15.85% (e) The percentage must be somewhere between 2.3% and 15.85%. (f) The percentage is smaller than 0.15% and can most likely be considered negligible. (g) The z -score of 115 seconds for the fifth grader is +1.25, and the z -score of 136 seconds for the third grader is +0.75. Relative to the different grades, we can say that the third grader is faster than the fifth grader. (h) The z -score of 107 seconds for the fifth grader is -0.75 , and the z -score of 107 seconds for the third grader is -2.875 . Relative to the different grades, we can say that the third grader is much faster than the fifth grader. (i) 142 seconds (j) 104 seconds
- 15-2** (a) Since considerably less than 1% of items in a normally distributed population are more than three standard deviations away from the mean, we would conclude that the yearly household incomes in the state very likely do not have a normal distribution. (b) Since we do not know that the incomes are normally distributed, we cannot be certain that practically all of the yearly household incomes are within three standard deviations of the mean.

Summary

Just as the distribution for a sample can be displayed with a histogram or a frequency polygon, the distribution for a very large population can be represented as a smooth curve called a *density curve*. The proportion of area under a density curve between any two values on the horizontal axis represents the relative frequency of items which fall between the two given values; we may also interpret this relative frequency as the probability that one randomly selected item falls between the two given values. The mean and standard deviation for a population are represented respectively by μ and σ , while the mean and standard deviation of a sample are represented respectively by \bar{x} and s .

The symmetric, bell-shaped distribution named the *normal distribution* plays an important role in statistics, because the sampling distribution of \bar{x} with simple random sampling becomes more like a normal distribution as the sample size n increases. A normal distribution can be described entirely in terms of the number of standard deviations that values are way from the mean. Well over half of the items in a normally distributed population are within one standard deviation of the mean; more specifically, 68.26% of the area under a normal density curve is between $\mu - \sigma$ and $\mu + \sigma$. Most the items in a normally distributed population are within two standard deviations of the mean; more specifically, 95.4% of the area under a normal density curve is between $\mu - 2\sigma$ and $\mu + 2\sigma$. Practically all of the items in a normally distributed population are within three standard deviations of the mean. Even though a normal density curve extends infinitely in both directions without ever touching the horizontal axis, the amount of area under the curve outside three standard deviations is negligible; more specifically, 99.7% of the area under a normal density curve is between $\mu - 3\sigma$ and $\mu + 3\sigma$, leaving only slightly more than one quarter of one percent of the area outside three standard deviations.

If x represents the value of a variable corresponding to a population of values with mean μ and standard deviation σ , then the *z-score* or *standard score* of x is the number of standard deviations (σ s) x is away from the mean (μ); specifically, the *z-score*, or *standard score*, of x is

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

The *z-score* is negative when x is less than the mean and is positive when x is larger than the mean. We can use *z-scores* with any distribution, but they are primarily used with regard to normal distributions, because of their importance in describing normal distributions. The value x can be called a *raw score*. We convert a *z-score* back into a *raw score* by using

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