10 The Normal Distribution

In Chapter 2 the concept of distribution was introduced. A distribution of scores refers to the relative frequency of the various scores. In this chapter the distribution that is commonly assumed in data analysis—the normal distribution—is discussed. Not only is the normal distribution used in data analysis, it also underlies various data transformations.

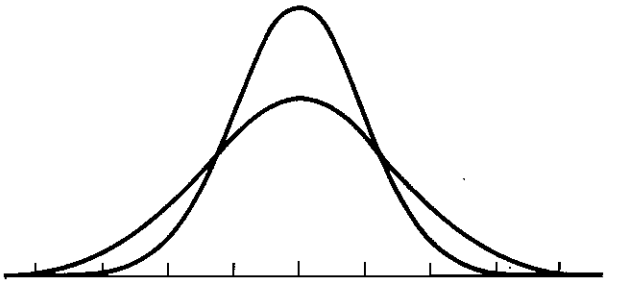
Properties of a Normal Distribution

The *normal distribution* is a symmetric, unimodal distribution that looks like a bell. It is a hypothetical distribution dreamed up by mathematicians that approximates the distribution of many naturally occurring variables. The upper and lower limits of the distribution are plus and minus infinity. Although all values are theoretically possible, very large or very small values are practically impossible. Because the normal distribution is a continuous distribution, any values, not just integers, are possible. Figure 10.1 shows two examples.

The normal distribution is not one distribution but actually a family of distributions. They differ only in their mean and variance. For instance, in Figure 10.1 both normal distributions have the same mean but one (the more peaked one) has less variance than the other. Although these distributions differ in their variance, their basic shape is exactly the same. If the mean and the variance of a normal distribution are known, then the exact shape of the distribution is known.

Although the two distributions in Figure 10.1 look quite different, they are both normal distributions. What is meant that they both have the same shape? Two distributions are said to have the same shape if there is a no-stretch transformation of one distribution that makes it possible to superimpose that distribution on the other.

FIGURE 10.1 **Examples of normal distributions.**

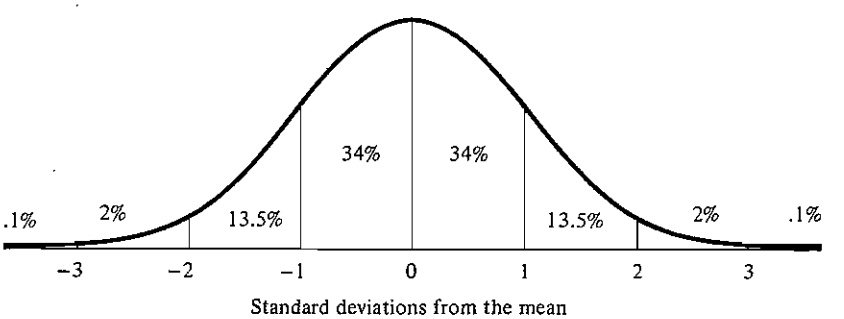


For the normal distribution the interval from the mean to one standard deviation above the mean contains about 34% of the scores, as is shown in Figure 10.2. In the interval from one standard deviation above the mean to two above the mean are about 13.5% of the scores. From two to three standard deviations are about 2% of the scores, and about .1% of the scores are greater than three standard deviations above the mean.

Stated alternatively, the interval from one standard deviation below the mean to one standard deviation above the mean contains about 68% of the scores. In the interval from two standard deviations below the mean to two standard deviations above the mean are about 95% of the scores. In the interval from three standard deviations below the mean to three above are about 99.7% of the objects. These facts hold for any normal distribution, regardless of the value of the mean and the variance.

Again, the normal distribution is a theoretical distribution dreamed up by mathematicians. Real numbers at best only roughly conform to its shape. For instance, the distribution of intelligence test scores is often assumed to be

FIGURE 10.2 **Probabilities for the normal distribution.**



normal. However, close inspection of the distribution reveals that there are more persons with very low intelligence than there would be if the distribution of intelligence were perfectly normal. Also, if the distribution were truly normal, an intelligence test score of –20 would be possible. The normal distribution is not "normal" in the sense that it is typical; actually it is quite atypical. Real, rather than theoretical, data are almost never exactly normally distributed. They may be close to normal but they are almost never exactly normal.

Even so, it is still reasonable to assume that the data are exactly normally distributed. There are four reasons for doing so. First, the normal distribution has some mathematically useful properties. For instance, $\bar{X}$ and s^2 are unrelated, which is not true of any other distribution. Another important fact is that if the variable X is normally distributed, then the distribution of $\bar{X}$ is also normal. If X has any other distribution, the mean of observations drawn from X has a different type of distribution. Data analysis is sufficiently complicated mathematically even if the normal distribution is assumed. If nonnormality is permitted, much of the algebra becomes quite difficult and, in certain cases, practically impossible.

Second, many variables that social and behavioral scientists study tend to be roughly, if not exactly, normally distributed. There is an important statistical theorem called the *central limit theorem* that explains this fact. The central limit theorem states that if a score is made up of a sum or average of n numbers, then as n gets larger the score tends to have a normal distribution even if the components are not normally distributed. Because many measurements in the social and behavioral sciences are often the sum of a large number of responses, it is all but inevitable that they have a distribution that approaches normality.

Third, even if the numbers are not normally distributed, in many cases they can be transformed to become "more normal" by the transformations discussed in Chapter 5. For instance, skewed distributions can be made more normal by applying one-stretch transformations. Thus, nonnormal data can be made "more normal" through data transformation.

Fourth, even if the data are not normally distributed, the errors resulting from nonnormality are often not that costly. For instance, using the statistical technique called analysis of variance (discussed in Chapters 14 and 15) with nonnormal data results in surprisingly few errors in most cases. The reason for this is the previously mentioned central limit theorem. So, the costs of falsely assuming normality are often only minimal.

A normal distribution has two parameters: its mean symbolized by μ and its variance symbolized by σ^2. Regardless of the shape of the distribution, the sample mean ($\bar{X}$) and the sample variance (s^2) are unbiased estimators of the population mean μ and the population variance σ^2, respectively. If the distribution is normal, $\bar{X}$ is a more efficient estimate of μ than either the median or the mode even though the median and the mode are unbiased

estimates of the population mean. The median and the mode have wider standard errors than the sample mean. In fact, it can be shown that $\bar{X}$ is the most efficient estimator of the population mean given a normal distribution.

Also, given normality, s^2 is a more efficient estimator of σ^2 than any other unbiased estimator of s^2. Thus, $\bar{X}$ and s^2 are preferred estimators of μ and σ^2 when the observations have a normal distribution. Because normal distributions are very commonly presumed in statistical work, both $\bar{X}$ and s^2 are the estimators of choice.

Standard Normal Distribution

In Chapter 5, the Z score transformation was presented. Its formula is

$$Z = \frac{X - \bar{X}}{s}$$

Because the statistics $\bar{X}$ and s are used to compute Z scores, this is a sample-based transformation. Imagine that X is normally distributed with a known population mean of μ and population variance of σ^2. The Z score in the population is computed by the formula

$$\frac{X - \mu}{\sigma}$$

Thus observations are adjusted by the parameters μ and σ and not by the sample-based statistics $\bar{X}$ and s. The above Z score has a normal distribution like X, but unlike X, the mean of the Z scores is always zero and the variance is always one. A normal distribution with a mean of zero and variance of one is called the *standard normal distribution* or *Z distribution*. Any normally distributed variable can become a standard normal variable by subtracting the population mean from each score and dividing this difference by the population standard deviation.

The standard normal distribution is a normal distribution in which the scores are expressed in standard deviation units. So if a variable has a Z distribution, a score of 1.5 indicates that the object is one-and-a-half standard deviations above the mean. A score of -1 indicates that the object is one standard deviation below the mean.

It is important to understand the difference between the Z transformation and the Z distribution. The Z transformation can be applied to any set of numbers regardless of their distribution. The Z transformation does not alter the basic shape of a distribution, only its mean and standard deviation. So a Z score transformation does not make a nonnormal distribution normal, as is sometimes mistakenly thought. However, when the Z transformation is applied in the population to a normally distributed variable, the result is the standard normal distribution. Though related, the Z transformation and the Z distribution are not the same.

It bears repeating that any normally distributed variable can be transformed into a variable with a standard normal distribution. Simply subtract the population mean and divide the difference by the population standard deviation. The resulting variable has a normal distribution with a mean of zero and a variance of one.

Determining Probabilities

The standard normal distribution is used to determine the probability of various types of events. To determine such probabilities Appendix C is used. In this appendix are listed the probabilities that a score is in the interval from zero to the value labeled Z in the table. For example, the probability of obtaining a Z score between 0.00 and 1.00 is .3413, given the mathematical properties of the standard normal distribution. It is then the case that the probability of being in the interval between the mean and one standard deviation above the mean is .3413 for any normally distributed variable. Thus, the table of probabilities for the standard normal distribution can be used to answer questions about any normal distribution for which the mean and variance are known.

To use Appendix C to determine the probability that a score will fall in the interval between zero and .50, first locate .50 in the left-hand column of the table. Then the number to the right, .1915 gives the probability.

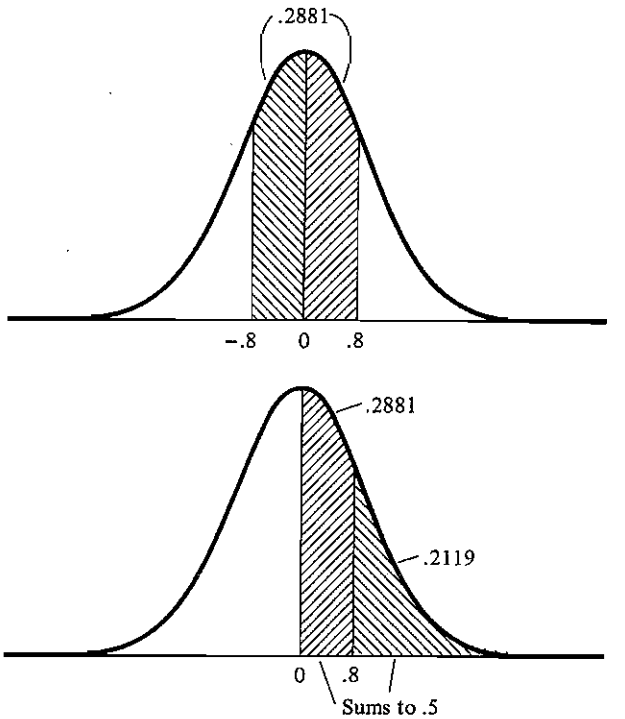
The probability that Z exactly equals any particular value is zero. Because a normal distribution is a continuous distribution, any value is possible—not just integers. For continuous distributions that can take on any value, only the probability of an interval is nonzero. The probability of being exactly at zero or at any particular value is zero.

Because the normal distribution is symmetric, it happens that the probability of something being in the interval between 0 and k is the same as it being in the interval between $-k$ and 0. So because the probability of being between 0 and .80 is .2881 (see Appendix C), then the probability of being between $-.80$ and 0 is also .2881.

Another useful fact concerns the probability of an event happening that is greater than some value, say k. This question can be reformulated as the difference between two probabilities. What is the probability of a score being greater than zero minus the probability of a score being between zero and k? The probability of a score being greater than zero is .5 because the normal distribution is symmetric and so one-half the scores must be above its mean of zero. As an example, what is the probability that a Z score is greater than .80? Because from Appendix C the probability of Z being between 0 and .80 is .2881, the probability of Z being greater than .80 is .5 minus .2881 or .2119. These facts are illustrated graphically in Figure 10.3.

To compute the probability of sampling within some interval of a variable that is normally distributed, the numbers are first transformed into Z scores. The question is reformulated so that it can be answered using the probabilities

FIGURE 10.3 **Facts about the Z distribution.**



given in Appendix C. Consider some examples using men's height as the variable, which is assumed to have a mean of 70 inches, a standard deviation of 3 inches, and a normal distribution (Stoudt, 1981).

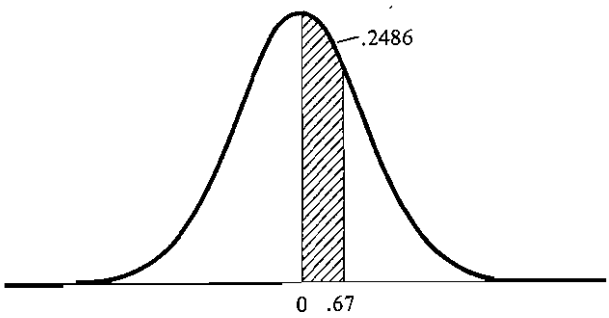
a. What is the probability that a man is between 70 and 72 inches tall? First, the numbers 70 and 72 are converted into Z scores. These Z scores are

$$\frac{70 - 70}{3} = 0$$

and

$$\frac{72 - 70}{3} = .67$$

The question now becomes what is the probability of obtaining a Z score between 0 and .67. The answer from Appendix C is .2486, as shown here graphically.



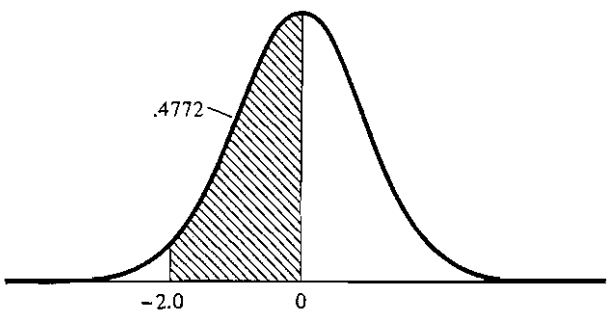
b. What is the probability that a man is between 64 and 70 inches tall? Both 64 and 70 are converted into Z scores:

$$\frac{64 - 70}{3} = -2.0$$

and

$$\frac{70 - 70}{3} = 0.0$$

The question becomes: What is the probability of a Z score being between –2.0 and 0? Because Z is symmetric, the question can be rephrased: What is the probability of sampling someone between 0 and 2.0? The answer from Appendix C is .4772, as shown graphically.



c. How likely is it that a man is between 65 and 72 inches tall? The numbers 65 and 72 are converted into Z scores:

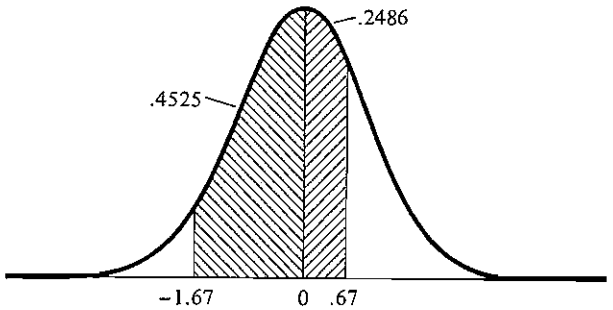
$$\frac{65 - 70}{3} = -1.67$$

and

$$\frac{72 - 70}{3} = .67$$

To answer this question, it is divided into two separate parts. The

probability of being between 0 and 1.67 is .4525 and between 0 and .67 is .2486. Both of the probabilities can be ascertained from Appendix C. So the probability of being between the interval –1.67 and .67 is .4525 + .2486, which equals .7009, as shown graphically.



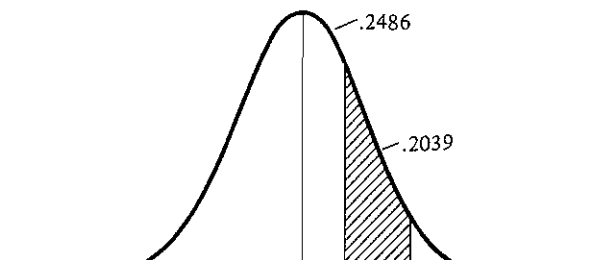
d. How likely is it that a man is between 72 and 75 inches tall? First, 72 and 75 are converted into Z scores:

$$\frac{72 - 70}{3} = .67$$

and

$$\frac{75 - 70}{3} = 1.67$$

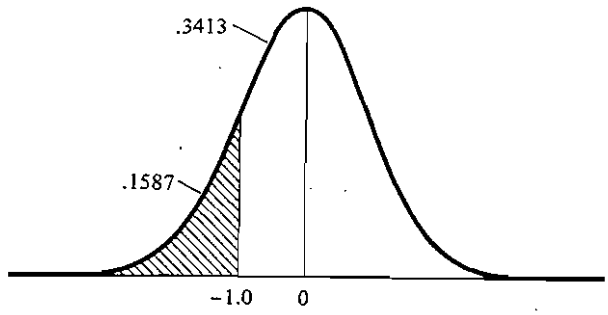
The question now is what is the probability of a Z being between .67 and 1.67. From Appendix C, the probability of Z being between 0 and .67 is .2486 and the probability of Z being between 0 and 1.67 is .4525. Thus, the probability of Z being between .67 and 1.67 is .4525 – .2486 = .2039, as shown graphically.



e. How likely is it for a man to be shorter than 67 inches? First, 67 is converted into a Z score:

$$\frac{67 - 70}{3} = -1.0$$

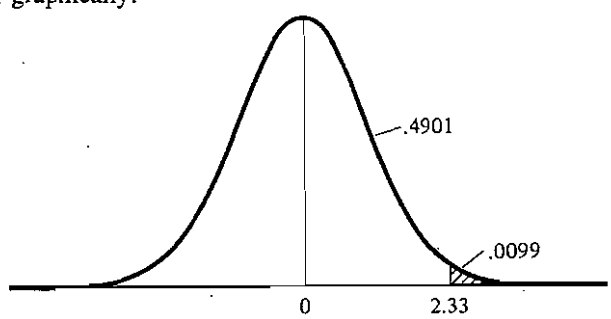
This is equivalent to asking the probability of a Z score greater than 1.0. The probability that Z is greater than zero is .5. From Appendix C, the probability that Z is the interval between zero and 1.0 is .3413. Thus, the probability that Z is greater than 1.0 is .5 – .3413 = .1587, as shown graphically.



f. What is the probability that a man will be 77 inches or taller? The score 77 is converted to a Z score:

$$\frac{77 - 70}{3} = 2.33$$

The question is: What is the probability that a Z score will be greater than 2.33? Because .50 is the probability of Z being greater than zero, and because .4901 is the probability of being between 0 and 2.33, the probability of Z being greater than 2.33 is .50 – .4901 = .0099, as shown graphically.



Determining Percentile Ranks

The standard normal distribution can also be used to determine a score's percentile rank. The percentile rank states the percentage of objects that the given object scores higher than. For instance, Scholastic Aptitude Test (SAT) scores are often expressed in terms of percentile ranks; for example, John has

an 87 percentile rank on verbal SAT. He scored higher than 87 percent of the people taking the test.

If one knows a person's score on a test, the test's mean and standard deviation, and one can assume that the variable is normally distributed, the person's percentile rank can be determined. The rule is simple: convert the score to Z and determine the probability value in Appendix C. If Z is positive then add .5 to probability and then multiply by 100 to get the percentile rank. If the Z is negative, the probability is subtracted from .5 and multiplied by 100.

The process can also be reversed. If the percentile rank is known, the Z score can be determined. One could use Appendix C, but it is simpler to use Appendix A. One first finds the percentile rank in the column denoted Proportion and then reads the Z score from the column labeled Probit.

The percentile ranks are used, in part, to determine the recommended daily allowances (RDA) of vitamin intake. Nutritionists survey the population of healthy individuals and calculate the mean and variance of intake for a particular vitamin. Using the mean and variance and assuming normality, the researchers calculate the value of a score with a 97.5 percentile rank. This score is used as the recommended daily allowance (RDA) for many vitamins.

Data Transformations

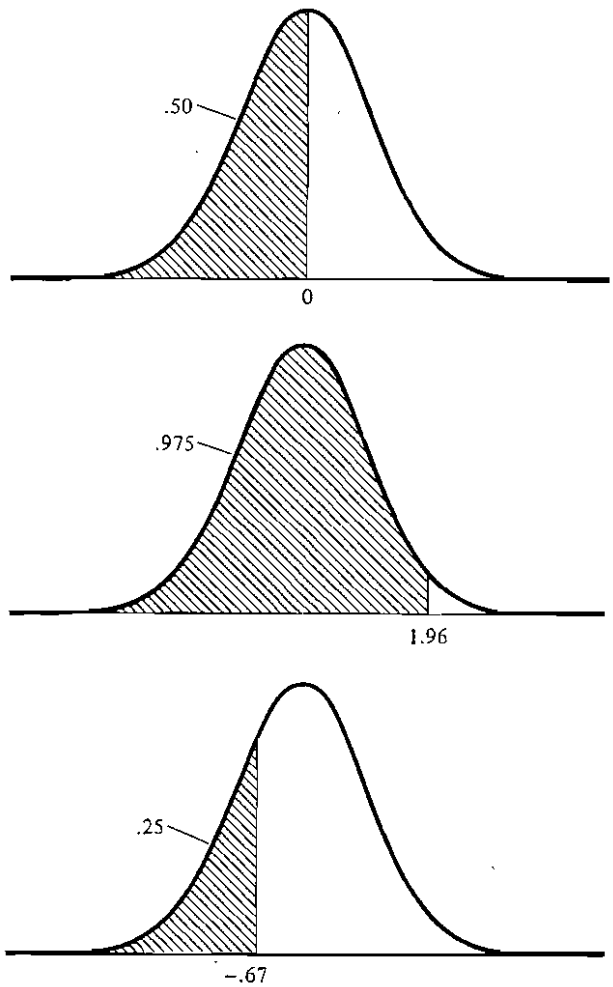
In Chapter 5 various data transformations are considered. Two of the transformations, the probit and percentile rank, described in that chapter can be clarified at this time.

Proportions

The probit transformation is used for proportions and percentages. It is a two-stretch transformation used to remove the lower limit of zero and the upper limit of 1.00 in proportions and 100 in percentages. This transformation is based on the standard normal curve. The probit of a proportion refers to the X axis of the normal curve for that proportion. So the probit transforms a proportion into a Z score.

The probit transformation is illustrated in Figure 10.4. In the top diagram the shaded area contains .50 of the distribution, and the reading on the X axis is 0.00. So the probit of .50 is zero. In the middle diagram the shaded area contains .975 of the distribution and the value on the X axis is 1.96. So the probit of .975 is 1.96. In the bottom diagram the shaded area contains .25 of the distribution and the value on the X axis is $-.67$. So the probit of .25 is $-.67$. (Sometimes a value of five is added to the probit to eliminate negative values.)

FIGURE 10.4 **Probit examples.**



Ranks

The normal distribution is also used to transform scores that have been rank ordered, but the underlying process that generated the scores is assumed to be normal. The transformation is called the *normalized ranks transformation*. It involves two steps. First, the ranks are transformed into percentile ranks by the formula presented in Chapter 5. The formula is

$$100\left[\frac{R - .5}{n}\right]$$

where R is the rank order of the score and n the sample size. Second, these percentile ranks are converted into Z scores using the probit transformation in Appendix A. The resulting values are the scores' normalized ranks.

The normalized ranks transformation can also be used to normalize the scores of any variable that is not normally distributed. First the scores are rank ordered from the smallest to largest. Then each score's percentile rank is computed. Finally, using the probit transformation in Appendix A, the percentile ranks are converted into Z values. A frequency distribution of the transformed scores is more nearly normally distributed than the untransformed scores.

The use of the normalizing transformation should be applied cautiously. If the obtained distribution appears implausible and there is good reason to believe that the variable is normally distributed, then a normalizing transformation may be helpful. It should not be routinely applied to nonnormal data, however.

As an example, consider the following sample.

15, 19, 21, 21, 34, 50, 52

These scores rank ordered are

1, 2, 3.5, 3.5, 5, 6, 7

Converting the scores to percentile ranks yields

7, 21, 43, 43, 64, 79, 93

Using the probit values in Appendix A, the normalized ranks are

−1.476, −.806, −.176, −.176, .358, .806, 1.476

Summary

The *normal distribution* is a symmetric, unimodal, bell-shaped distribution. Because of its relative mathematical simplicity, it is commonly used in statistical work. Also because of the central limit theorem, many statistics have approximately a normal distribution. The *central limit theorem* states that the sum of numbers tends to be normally distributed as more numbers are summed.

The *standard normal distribution* is a normal distribution with a mean of zero and a variance of one. The standard normal is often referred to as the Z *distribution*. This distribution can be used to answer questions about the likelihood of certain types of events as well as to compute percentile ranks.

The normal distribution is used for various data transformations. Proportions can be transformed by the *probit transformation* and ranks by the *normalized ranks transformation*.

Problems

1. Find the following probabilities from the Z distribution.
   a. between 0.0 and .77
   b. between −.11 and 0.0
   c. between .33 and 1.35
   d. greater than .72
   e. between 0.0 and 2.04
   f. between −1.22 and 0.0
   g. between −1.48 and −.99
   h. less than −.38
   i. between −.46 and 1.13
2. Let IQ be a normally distributed variable with a mean of 100 and a standard deviation of 15. Find the probability that someone's IQ is
   a. between 100.0 and 115.0
   b. between 110.0 and 120.0
   c. less than 95
   d. between 90.0 and 100.0
   e. between 90.0 and 95.0
   f. less than 123
3. If X is a normally distributed variable with a mean of 12 and a variance of 16 what is the probability of the following sets of events?
   a. X between 10.0 and 12.0
   b. X less than 11.0
   c. X greater than 12.5
   d. X between 11.0 and 14.0
4. Find the probit transformations of the following probabilities.
   a. .66 b. .10 c. .55 d. .34
5. Convert the following rank-ordered scores into normalized ranks.

   1, 2, 3, 4, 5, 6, 7, 8, and 9

6. Answer the following questions, assuming that the distribution is normal.
   a. How likely is it for someone to score at least 1.5 standard deviations above the mean or more?

b. How likely is it for someone to score lower than 1.75 standard deviations below the mean?
c. What is the probability of someone scoring between .5 standard deviations below the mean to .5 standard deviations above the mean?

7. Below are the rank-order scores of ten cities in the United States as rated by Rand McNally (Boyer & Savageau, 1985) on three dimensions.

City	*Transportation*	*Economics*
Atlanta	2	5
Boston	6	4
Chicago	5	10
Cincinnati	8	8
Dallas	7	1
Denver	4	2
New York	1	7
Phoenix	10	3
Pittsburgh	9	9
San Francisco	3	6

Convert the ranks to normalized ranks and average the normalized ranks for each city. Compare these averages to the means of the original ranks for each city.

8. Given that X is normally distributed with a mean of μ and a variance of σ^2, it is true that $\bar{X}$ has a mean of μ and a variance of σ^2/n. Given this fact, if X has a mean of 50 and a variance of 81, what is the probability that $\bar{X}$ will be between 51 and 49 if n is 36?

9. Given that scores on the Scholastic Aptitude verbal test have a mean of 500 and a standard deviation of 100 and a normal distribution, what percentage of the population is outscored if the following scores are obtained?

a. 600 b. 700 c. 500
d. 750 e. 450 f. 350

10. If the probability that of being five units or more above the population mean is .25 and the distribution is normal, what is the standard deviation of the variable?

11. Given that weight for females 18–24 is normally distributed with a mean of 132 and a standard deviation of 27 (Stoudt, 1981), compute percentile ranks for the following weights:

a. 132 b. 150 c. 95
d. 139 e. 180 f. 100

12. For the following percentile ranks, determine the corresponding Z values:

    a. 60 b. 31 c. 29
    d. 48 e. 79 f. 93

13. Explain the following statement: The Z transformation does not make a distribution normal, but the normalized ranks transformation does.

14. Convert the following scores into normalized ranks.

    418, 423, 425, 425, 430, 435, 435, 440, 441