Calculating Standard Deviation for a Sample

You will recall from chapter 1 that most research is performed on samples taken from populations. The sample is always smaller than the population and should be selected randomly so that the statistics calculated from the sample are representative of the corresponding parameters in the population. The researcher assumes that conclusions based on the sample are applicable to the population from which the sample was drawn.

Samples rarely contain the extreme values that are found in the population. For example, if we randomly sampled the weights of 100 men in a university with 15,000 male students, it is not likely that anyone in the sample would weigh 350 pounds or 100 pounds, although there may be students in the population who weigh these amounts. The variability of the sample is almost never as large as the variability of the population.

When standard deviation is calculated from a sample and then used to estimate the standard deviation of the population, a correction factor must be applied to the equation so that the estimate of the population is not biased by a small sample. Without this correction factor, an estimate of a population standard deviation based on a sample would be erroneously small.

The equations presented previously in this chapter for calculating standard deviation are based on the assumption that an entire population has been measured. If these equations were applied to samples, an error when generalizing from a sample to a population would occur. The correction to the equations is based on the concept of degrees of freedom.

The **degrees of freedom** are the number of values in a data set that are free to vary. If no restrictions are placed on the data, then all values are free to vary; that is, they may take on any value, limited only by the precision of the measuring instrument and the actual values in the population. But when we take a sample and make the assumption that the sample represents the population, a restriction is placed on the data in the sample.

When the sample mean is assumed to be identical to the population mean (i.e., the sample mean is established), the sum of the sample data is also set because the mean equals the sum divided by N. This limits the numerical value of the last data point in the sample to a number that will create a sample statistic theoretically equal to the population parameter (even though the population value is unknown).

Assume that the mean of four values must be 5.0. Therefore, the sum of the four numbers must equal 20. Let the values 2, 3, and 7 be the first three numbers. What must the fourth number equal to bring the sum to 20? It must be 8. The last number is not free to vary; it is limited to only one value, that which will create a sum of 20. Therefore, this example has 3 degrees of freedom. Three of the four numbers can assume any value, but the last one must be whatever it takes to make the sum equal 20. The degrees of freedom for a single data set that is a sample representing a population are always $N - 1$. In this example, $df = N - 1 = 4 - 1 = 3$.

This correction factor, when applied to the definition formula for standard deviation (equation 5.03) reduces the denominator by 1. The standard deviation for a sample is

$$SD = \sqrt{\frac{\Sigma d^2}{N-1}}. \tag{5.07}$$

Note that the symbol for standard deviation in equation 5.07 is SD rather than σ. In this book, we use SD to represent the standard deviation for a sample and σ to represent the standard deviation for a population. (In some statistics books, standard deviation may be represented by S or s.)

Like equation 5.03, equation 5.07 is difficult to apply when N is large and the mean is not an integer, and it is almost never used to calculate standard deviations. But it may be modified so that deviation scores are not required. The modified formula is

$$SD = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N-1}}. \tag{5.08}$$

Equation 5.08 is commonly used to calculate standard deviation from samples arranged in a rank order distribution. If we apply equation 5.08 to the data from table 5.4, the result is

$$SD = \sqrt{\frac{19{,}275 - \frac{(337)^2}{6}}{6-1}} = 8.3.$$

The standard deviation calculated by equation 5.08 yields a slightly larger answer (SD = 8.3) than does equation 5.05 (σ = 7.4). The difference represents the correction for degrees of freedom. If the data in table 5.4 are assumed to represent a sample from a larger population, the estimate of the population standard deviation would be 8.3.

For use on simple frequency data, equation 5.08 may be modified by adding f. The resulting equation is

$$SD = \sqrt{\frac{\Sigma fX^2 - \frac{(\Sigma fX)^2}{N}}{N-1}}. \tag{5.09}$$

In the Olympics, all performances (even the last-place finishers) are superior to those of non-Olympic athletes. If we graded scores in a high school physical education class by comparing them with Olympic performances, all the students would fail.

How good is a 4-meter long jump in a high school class? Without additional information, we can't be sure. But if we know that the average of all long jumpers in the class is 5 meters, then we know that this score is below the mean, but how far below? To evaluate a single raw score, we must compare it to a scale that has known central tendency and variability. Scores from such scales are called standard scores. Standard scores provide information that helps us evaluate a given raw score.

When studying the concept of variability, especially while learning about standard deviation, students often ask, What does standard deviation mean? They know how to calculate it, but they do not understand its meaning or its value. The answer lies in an understanding of the unique characteristics of the normal curve.

Observed over a sufficient number of cases, many variables will assume a normal distribution. This is fortunate for the statistician because the characteristics of the normal curve are well known. If the data are from an interval or a ratio scale, if enough cases have been measured, and if the curve is normal or nearly normal, the characteristics of normality may be applied to the data.

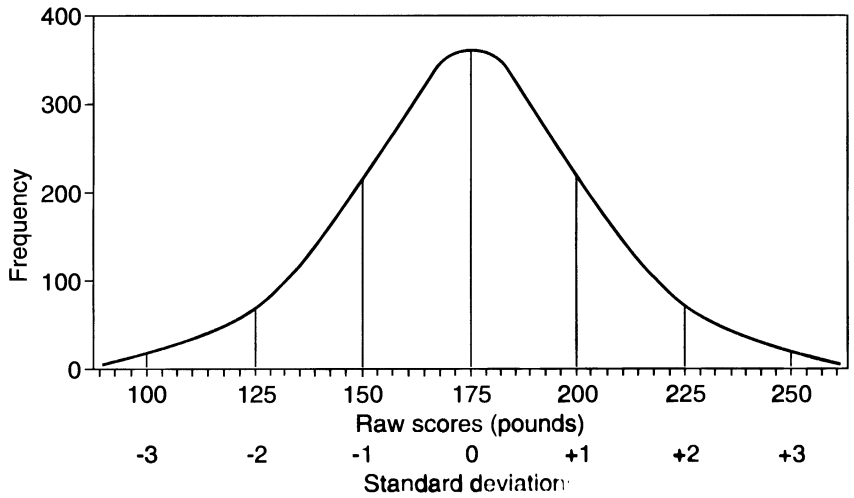

Figure 6.1 Relationship between raw scores and *Z* scores on a normal curve.

Figure 6.1 shows a frequency polygon of data on the weights of a population ($N > 1{,}000$) of college-age males. The mean weight is 175 pounds, and the standard deviation is 25 pounds. The graph shows that most subjects weigh between 150 and 200 pounds ($\bar{X} \pm 1\sigma$) and that the highest frequency is at the mean (175 pounds). A few subjects weigh less than 100 pounds, and a few weigh more than 250, but not many.

Most subjects' weights are near the mean. As values progress farther from the mean in either direction, fewer and fewer cases are represented. The standard deviation units are distributed equally above and below the mean, and the majority of the cases fall between the mean and $\pm 1\sigma$, or between 150 and 200 pounds.

Z Scores

A ***Z* score** is a raw score expressed in standard deviation units. If the standard deviation of the scores in figure 6.1 is 25, then 1 standard deviation unit is equivalent to 25 pounds on the raw score scale. A score of 200 lies 25 raw score units, or 1 standard deviation unit, above the mean. The raw score of 200 is equivalent to a *Z* score of +1. Likewise, a raw score of 150 (25 raw units, or 1 standard deviation unit, below the mean) has a *Z* score of –1.

When the mean and the standard deviation of any set of normal data are known, *Z* can be calculated for any raw score (X) by using the following formula:

$$Z = \frac{X - \bar{X}}{\sigma}. \qquad (6.01)$$

NOTE: If the data represent a sample, use SD in the denominator.

Using this formula, we can calculate that a male student who weighs 200 pounds has a *Z* score of +1:

$$Z = \frac{200 - 175}{25} = +1.$$

When the raw score is less than the mean, the *Z* score is negative. If the raw score is 165,

$$Z = \frac{165 - 175}{25} = -0.4.$$

A unique characteristic of the normal curve is that the percentage of area under the curve between the mean and any *Z* score is known and constant. When the *Z* score has been calculated, the percentage of the area (which is the same as the percentage of raw scores) between the mean and the *Z* score can be determined.

In any normal curve, 34.13% of the scores lie between the mean and one *Z* score in either the positive or negative direction. Therefore, when we say that most of the population falls between the mean and ±1 *Z* score, we are really saying that 68.26% ($2 \times 34.13 = 68.26$), or about two thirds, of the population falls between

these two limits. This is true for any variable on any data provided that the distribution is normal. Figure 6.2 demonstrates this concept.

Table A.1 in appendix A may be used to determine the percentage of scores that falls between the mean and any given *Z* score. The numbers on the left and right sides of the table represent *Z* scores to the nearest tenth of a point. The values across the top represent *Z* scores to the nearest hundredth.

To determine the percentage of scores that falls between the mean and ±1.00 *Z* score we proceed down the left-hand column to the value 1.0 and move across the row to the .00 column; the value in that column is 34.13. This value is the percentage of raw scores that falls between the mean and either +1.00 *Z* or –1.00 *Z*. For ±2.00 *Z* the value in table A.1 is 47.72 (which is equal to 34.13 + 13.59, the percentages of the area from 0 to +1 and from +1 to +2 in figure 6.2). This indicates that 47.72% of the population of scores lies between the mean and +2.00 *Z* scores, or between the mean and –2.00 *Z* scores.

Doubling that number (47.72 × 2 = 95.44) tells us that slightly more than 95% of all raw scores lies between the mean and ±2 *Z*. The figure for 3 standard deviations is 99.74% (49.87 × 2 = 99.74). This confirms that most raw scores fall within ±3 standard deviations of the mean (see figure 6.2).

Because the normal curve is bilaterally symmetrical, table A.1 in appendix A provides only the values for one-half of the curve; the values are the same for positive and negative *Z* scores. The percentage corresponding to a *Z* score of –1.78 is found by proceeding down the left-hand column to 1.7, then across to the .08 column, where the value 46.25 is read. This is interpreted to mean that 46.25% of the population of scores lies between the mean and 1.78 *Z* scores in either direction.

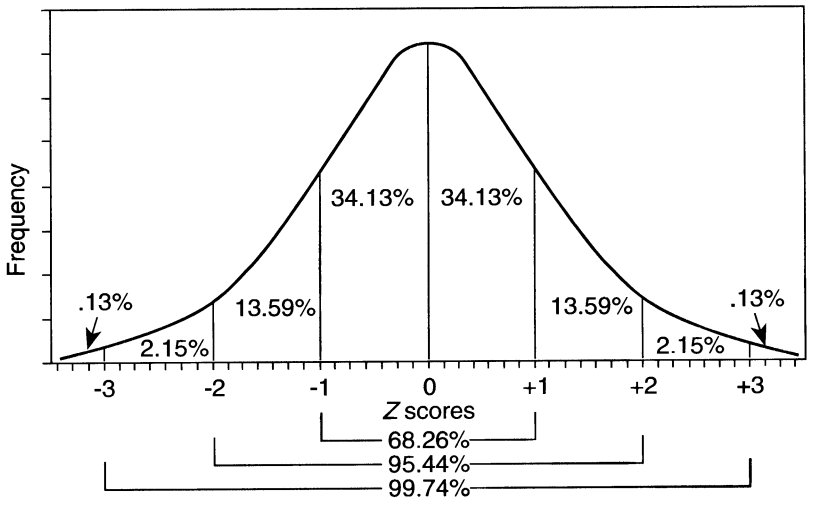

Figure 6.2 Percentage of area under the normal curve for selected *Z* score values.

Table A.1 may also be used for the opposite procedure, to find the *Z* score that corresponds to a given percentage of the population. If we want to know the *Z* value that represents 30% of the area under the curve, we look for the figure of 30.00 in the body of the table. This exact number cannot be found, so we find the closest value to 30.00 (29.95). This value corresponds to a *Z* score of ±0.84. So to be 30 percentage points above or below the mean, a person must have a raw score equal to about ±0.84 *Z* scores.

Converting Z Scores to Percentile Scores

Once we know the *Z* score for a raw score, we can also determine the percentile value of that score by looking in table A.1 in appendix A. Because the mean, or a *Z* score of 0.00, represents the 50th percentile, any figure read from the table for a positive *Z* score is added to 50 to determine the percentile of that score. A *Z* score of +1.24 has a corresponding table value of 39.25. Therefore the raw score equivalent to a *Z* score of +1.24 has a percentile value of 89.25 (50 + 39.25 = 89.25). If the *Z* score is negative (–1.24), then the value from the table is subtracted from 50. This results in a percentile score of 10.75 (50 – 39.25 = 10.75).

Just as any raw score has a corresponding *Z* score, so each *Z* score has a corresponding percentile score. Table A.1 may be used to convert scores from *Z* to percentiles, or vice versa. It's important to remember that a positive *Z* score must be added to 50, and a negative one must be subtracted from 50, to determine the percentile value of the raw score. The values in table A.1 represent only half of the curve; they should be interpreted as the distance from the middle of the curve toward either end.

Standard Scores

Raw scores are the direct result of measurement, usually in units of distance, time, force, or frequency. Raw scores from more than one variable will have different units of measurement, different mean values, and different variability. A standard score is derived from raw data and has a known central tendency and variability. Raw scores from multivariate data can be directly compared only after they are converted to the same standard score base.

It is not logical to compare 50 sit-ups in 1 minute with a mile-run time of 4:36.3. Which performance is better, 20 feet in the long jump or 10.5 seconds in the 100-meter dash? We cannot answer such questions using raw data alone, because the values are based on different units of measurement. To make an appropriate comparison, we must first convert raw scores to one of the four standard scores: percentiles, *Z* scores, *T* scores, or stanines.

Percentiles

In chapter 3, we discussed the percentile. This type of standard score has several advantages. Percentiles have 50% as their central tendency and 0% to 100% as their range. They also have known quartile and decile divisions.

We can easily compare different types of raw scores when we convert them to percentiles. If a 5'5" high jump represents the 75th percentile, and a time of 11.5 seconds in the 100-meter dash represents the 80th percentile, then the running score is better than the jumping score. These scores are now directly comparable because they have both been converted to the same base, or standard.

Z Scores

The calculation of Z scores was explained earlier in this chapter. We will now discuss how to use Z scores as standard scores. Z scores have a known central tendency ($\bar{Z} = 0$) and known variability ($\sigma = 1.0$). When raw scores are converted to Z scores, two or more sets of data may be directly compared. For example, which score is better, a Z score on a long jump test of -1.3 or a Z score of -0.50 on a gymnastics floor exercise test? The gymnastics score is better, because a Z score of -0.50 is higher (thus better) than -1.3.

We can confirm this by using table A.1 in appendix A to convert both scores to percentiles: -1.3 Z equals a percentile score of 9.68 (50 – 40.32 = 9.68), and -0.5 Z equals a percentile score of 30.85 (50 – 19.15 = 30.85). The student who received these scores is not a whiz at either the long jump or floor exercise, but it is safe to say that the student is a better performer in gymnastics. Figure 6.3 presents a graphic representation of the two scores.

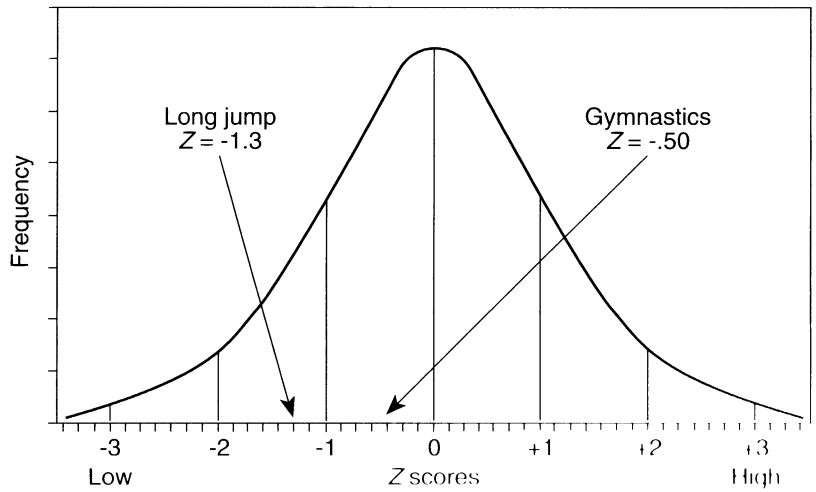

Figure 6.3 Comparison of Z score value on the normal curve

T Scores

A third standard score, called the T score, is often used to report norms in educational settings, such as those on national fitness or skill tests. By convention and definition, a ***T* score** has a mean of 50 and a standard deviation of 10. A T score of 60 is 1 standard deviation above the mean, and a T score of 30 is 2 standard deviations below the mean. Before T can be calculated, the corresponding Z score must be known. Then the formula for converting from Z to T scores is

$$T = 10Z + 50. \tag{6.02}$$

The T equivalent for a Z score of +1.5 is 65.0 ($T = 10 \times 1.5 + 50 = 65.0$). When the Z score is negative, the T score will be less than 50 (remember that Z scores have a mean of 0). In the previous example of gymnastics and long jump scores, the T score for a Z score of -0.50 in gymnastics is calculated as follows: $T = 10 \times (-0.50) + 50 = 45.0$.

The T scale was created because the lay public has difficulty understanding Z scores. It is not common to think of scores with 0 as the mean. Most people consider 0 to be nothing and prefer to have 50 as the middle and a range from 0 to 100. This is why percentiles are so widely used; they are easy to understand.

T scores have a mean of 50 and a range from 0 to 100, but it is very unlikely that a T score would be less than 20 or greater than 80 because these figures represent Z scores of -3 and +3, respectively. As figure 6.2 shows, only 2×0.13, or 0.26%, of the population lies beyond the ± 3 limit on the Z scale. Indeed, a T score of 100 would represent a Z score of +5 and a percentile of 99.99999, which is a rather unlikely occurrence.

Figure 6.4 shows the relationships among Z scores, percentiles, and T scores.

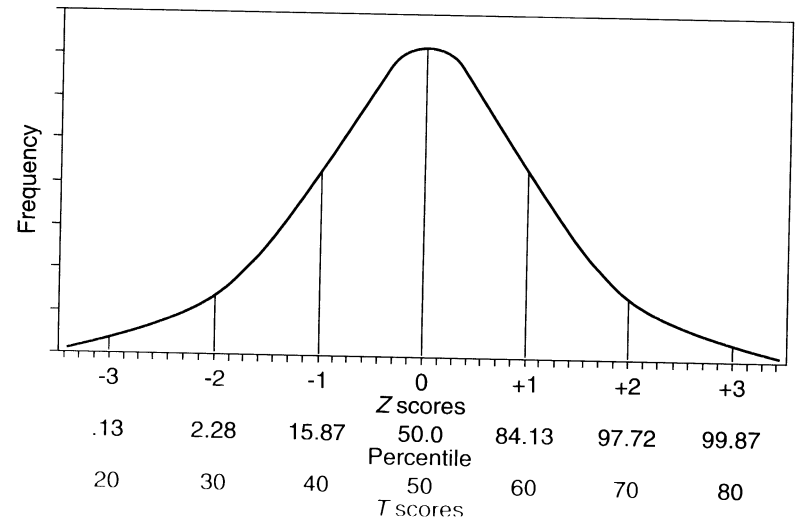

Figure 6.4 Relationships among Z scores, percentiles, and T scores on the normal curve

Stanines

Like the *T* scale, the **stanine** scale (a derivation of the words standard nine) is commonly used for reporting the results of educational tests. Parents who inquire about their children's scores on standardized tests may find the results presented in stanines. For example, a student may score at the 7th stanine in math, the 4th stanine in reading, and the 5th stanine in verbal skills. Physical education teachers need to understand stanines so they can read the reports found in the student files.

How are stanine scores interpreted? As figure 6.5 shows, in the stanine scale, the standard normal curve is divided into nine sections with 5 as the middle score, 1 as the lowest score, and 9 as the highest score.

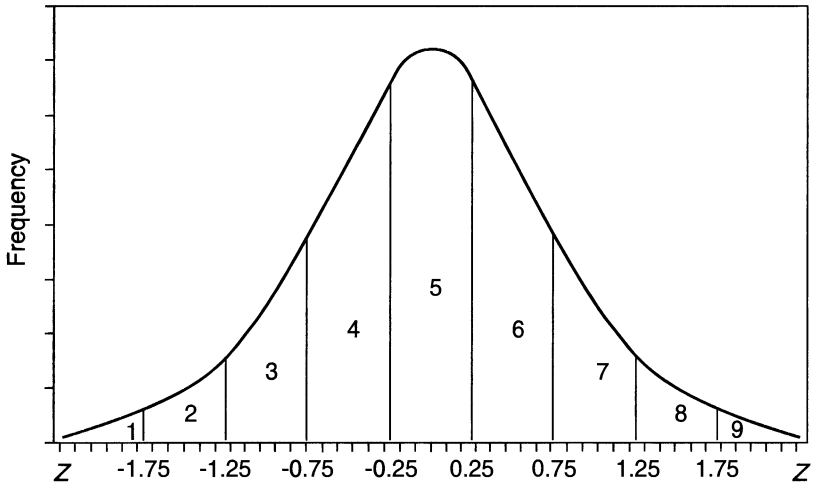

Figure 6.5 Stanine distribution on the normal curve.

To calculate a stanine score, we need to know the raw score, the mean, and the standard deviation. From this information, we calculate the *Z* score. Once we have the *Z* score, we can find the stanine score directly from figure 6.5.

Each section on the stanine curve is one-half *Z* score wide except for stanines 1 and 9. The center section (stanine 5) ranges from −0.25 *Z* to +0.25 Z. The other sections continue to the left or right in 0.5 *Z* intervals.

Scores that fall exactly on a dividing line between 2 stanines are usually given the higher value. Thus, a *Z* score of +0.75 is considered to be in the 7th stanine, and a *Z* score of −1.25 is in the 3rd stanine.

Stanine scores do not represent an exact raw score. Rather they represent a range, or section of the curve, into which the raw or *Z* score falls. In this way, stanines are similar to quartiles or deciles. When only the stanine is known, it is impossible to tell exactly where a raw score falls within the stanine. Only the section of the curve that best represents the score is known.

A student who scored at the 7th stanine on pull-ups, the 4th stanine on the mile run, and the 5th stanine on the sit-and-reach test of flexibility is considerably above the mean in strength, slightly below the mean in aerobic capacity, and very close to the mean in flexibility.

Predicting Population Parameters Using Statistical Inference

In many sciences, little research is conducted on entire populations. Often the population is so large that it would be impossible to measure each member. In such cases, the researcher takes a sample of the population and assumes that the sample represents the population and that the characteristics of the sample statistics are indicative of the population parameters. For this assumption to be valid, the sample must be randomly selected. For more information on sampling selection, see chapter 1. The process of estimating population parameters based on sample statistics is called **inferential statistics**.

Earlier we discussed the problem of determining the mean weight of all men at a university. If there were 100 or fewer men in the population, we could measure all of them, and a sample would not be needed. But if the population of all men in the university were 15,000, it would be too time-consuming to measure all of them. So we would take a random sample to *estimate* the mean of the population.

The sample size is limited by such factors as time restraints, lack of finances, and facilities and equipment. If we wanted to measure height or weight, a large sample could be collected because it is easy to measure these variables. But if we were interested in hydrostatically measured body composition or $\dot{V}O_2max$ on a treadmill or bicycle ergometer, perhaps only a few subjects could be measured.

Let us assume that a sample size of 50 men is desired. If the administration permitted us access to the student files, a computer could select 50 males from the files in a completely random fashion. These randomly selected subjects could then be invited to participate in the study and could be measured by appointment. (It would be wise to invite more than 50 subjects because a few may not participate.) The mean weight of this sample could then be used to represent the mean weight of all men at the university.

Estimating Sampling Error

Sampling error refers to the amount of error in the estimate of a population parameter that is based on a sample statistic. Even if the sample is randomly drawn,

it is unlikely that the mean of the sample will be identical to the mean of the population. Also, the true population mean is never known exactly because all members of the population are never measured. Consequently, we need a way to determine how accurate the sample mean is and what the odds are that it is deviant from the population mean by a given amount.

The **standard error of the mean** is a numeric value that indicates the amount of error that may occur when a random sample mean is used as a predictor of the mean of the population from which the sample was drawn. By accepting this error, we admit that the exact population mean can never be known (unless every member of the population is measured). We can only know the sample mean and how much error it is estimated to have.

The prediction of the population mean is always an educated guess and is accompanied by a probability statement. That is, the population mean is assumed to exist between some set limits, and the chance of this assumption being correct is stated as odds such as 90 to 10 ($p = .10$), 95 to 5 ($p = .05$), or 99 to 1 ($p = .01$).

Let us use the example of estimating mean weight of men at a university to explain the theory behind this technique. We assume that the range of weight of all men in the population is about 100 to 250 pounds and that the mean is about 175 pounds (a value we can never know precisely unless we measure all 15,000 men).

Suppose a large number (theoretically an infinite number) of random samples is taken with 50 subjects in each sample. After each sample is taken, the subjects are returned to the population pool so that they have an equal chance of being chosen again in a subsequent sample. Most samples have means between 165 and 185 pounds and ranges of about 125 to 225 pounds. The range of the population (which we estimated to be 100 to 250 pounds) will be larger than the range of any one sample because it is unlikely that the extremes of a population of 15,000 will be randomly selected into a sample of 50.

If a true random sample of sufficient size is taken each time, the sample mean will not vary greatly from the actual population mean. It is unlikely that a random sample of $N = 50$ would have a mean value near one of the extremes of the population, because for this to happen, all subjects in the sample would have to be from one extreme of the population. But this unusual occurrence is more likely to happen if the sample size is small (e.g., 5). To avoid this potential error, we make our samples as large as possible within the limits of our resources. *The probability of obtaining a biased sample increases as the sample size decreases. The larger the sample, the smaller the error in predicting the population mean.*

After all of the samples have been taken and the mean of each calculated, the sample means could be arranged into a graph that would approximate a normal curve. The means of this series of random samples would be normally distributed, even if the population from which they were drawn is not normal. This concept is known as the **central limit theorem.**

Each sample has its own mean and standard deviation, and the total group of sample means also has a mean (the mean of the means) and a standard deviation (the standard deviation of the means). This value, the standard deviation of the means, is called the standard error of the mean.

Because the sample means form a normal distribution, all of the characteristics of normality apply to the curve of the sample means. The mean of the sample means becomes the best estimate of the true population mean because it represents the results of a large number of randomly drawn samples of 50 subjects each.

Let us assume that the mean of the sample means is 175 pounds, and the standard deviation of the sample means, or the standard error of the mean, is 3.5. Applying our knowledge about the relationship of percentile points to Z scores on a normal curve (see figures 6.1 and 6.2), and assuming that the mean of the sample means is the best estimate of the population mean, we can state that the population mean probably lies somewhere between $175 \pm 1 \times 3.5$ pounds (171.5 to 178.5).

Recall that approximately 68% of the area under any normal curve lies between ± 1 standard deviation (see figure 6.2), and that the remaining 32% lies outside these limits (16% on each end). The mean of the means (175) becomes the estimate of the population mean, and there is a slightly better than 2-to-1 chance (actually it is 68 to 32) that the estimate of $175 \pm 1 \times 3.5$ pounds is correct.

If we widen the population estimate (making it less precise but more encompassing) to ± 2 standard deviations ($175 \pm 2 \times 3.5$) and estimate that the population mean lies somewhere between 168 and 182, we increase the odds of being correct to better than 95 to 5 (95.44% of the area under the normal curve lies between ± 2 standard deviations of the mean; see figure 6.2). And if we make the estimate accurate to ± 3 standard deviations ($175 \pm 3 \times 3.5$) and estimate that the population mean is between 164.5 and 185.5 pounds, then the odds that our estimate is correct are better than 99 to 1.

Using this technique, we can make probability statements about the population mean with various degrees of accuracy. *The more precise, or narrow, the estimate, the lower the odds of being correct. As the estimate becomes more general, or broad, the odds of being correct improve.*

The process just described is intended to show the theory behind the concept of the standard error of the mean. In practice, it may be as difficult to measure a large number of samples of 50 subjects each as it would be to measure all 15,000 males in the population. Fortunately, an equation has been derived that will estimate the standard deviation of a series of theoretical sample means based on only one random sample.

The standard error of the mean estimated from one random sample is denoted by SE_M. This value is algebraically demonstrated by equation 6.03:

$$SE_M = \frac{SD}{\sqrt{N}}, \tag{6.03}$$

where SD is the standard deviation of the sample and N is the sample size. (Refer to equation 5.08 for the formula for SD.)

Using only one randomly drawn sample and equation 6.03, we can calculate the odds that the population mean lies within certain limits. In the example we have been discussing, suppose we take one random sample of 50 subjects with a mean of 175 pounds and standard deviation of 25 pounds. Then we can calculate that SE_M is equal to 3.5:

$$SE_M = \frac{25}{\sqrt{50}} = 3.5.$$

This value is actually a standard deviation on a normal curve; therefore, it is equivalent to a Z score of ±1.0. We may infer from this calculation that the mean of the population from which this sample was drawn has a 68% chance of being within the limits of 175 ±1 (3.5) pounds. The process of inference is represented by the following equation:

$$\mu = \bar{X} \pm 1(SE_M),$$

where $\bar{X}$ represents the sample mean and μ (mu in the Greek alphabet) represents the population mean. In our example, this equation would indicate μ is 175 ± 1(3.5), or somewhere between 171.5 and 178.5. This is sometimes written as $171.5 \leq \mu \leq 178.5$. A similar formula for calculating the standard error of a proportion is available when data are presented as percentiles (see discussion on the t test for proportions in chapter 8).

Levels of Confidence and Probability of Error

A **level of confidence** (LOC) is a percentage figure that establishes the probability that a statement is correct. It is based on the characteristics of the normal curve. In the previous example, the estimate of the population mean (μ) is accurate at the 68% level of confidence because we included one SE_M (i.e., 1Z) above and one SE_M below the predicted population mean.

But if there is a 68% chance of being correct, there is also a 32% chance of being incorrect. This is referred to as the **probability of error** and is written $p < .32$ (the probability of error is less than .32). The area under the normal curve that represents the probability of error is called **alpha** (α). Alpha is the level of chance occurrence. In statistics, this is sometimes called the error factor (i.e., the probability of being wrong because of chance occurrences that are not controlled). Alpha is directly related to Z because it is the area under the normal curve that extends beyond a given Z value.

Remember that SE_M is a standard deviation on a normal curve. By including 2 standard errors above and below μ (175 ±2 × 3.5, or 168 to 182 pounds), we increase our level of confidence from 68% to better than 95% and drop the error factor from 32% ($p = .32$) to about 5% ($p = .05$).

To be completely accurate when we use the 95% level of confidence, or $p = .05$, we should not go quite as far as 2 Z scores away from the mean. In table A.1 in appendix A, the value in the center of the table that represents the 95% confidence interval is 47.50 (95/2 = 47.50, because table A.1 represents only half of the curve). This corresponds to a Z score of 1.96. The value 1.96 is the number of Z scores above and below the sample mean that accurately represents the 95% level of confidence, or $p = .05$. The correct estimate of μ at $p = .05$ is 175 ±1.96 × 3.5, or 175 ± 6.9 pounds (168.1 to 181.9).

A similar calculation could be made for the 99% level of confidence by looking in table A.1 to find the value that reads 49.5 (99/2 = 49.5). This exact value is not found in the table. Because 49.5 is halfway between 49.49 and 49.51 in the table, we choose the higher value (49.51), which gives us slightly better odds. The Z score correlate of 49.51 is 2.58. To achieve the 99% level of confidence, we multiply SE_M by ±2.58. The estimate of the population mean at the 99% level of confidence ($p = .01$) is 175 ±2.58 × 3.5 or 175 ± 9.0 pounds. This may be expressed as $166 \leq \mu \leq 184$, $p = .01$.

Likewise, we could establish the 90% level of confidence by looking up 45% (90/2 = 45) in table A.1. The Z score correlate of 45% is 1.65, so μ is 175 ±1.65 × 3.5, or 175 ± 5.8 pounds, $p = .10$.

The level of confidence (chances of being correct) and probability of error (chances of being incorrect) always add to 100%, but by tradition, the level of confidence is reported as a percent and the probability of error (p) is reported as a decimal. The Z values to determine p at the most common levels of confidence are listed in table 6.1. Other values may be determined for any level of confidence by referring to table A.1 in appendix A.

The generalized equation for determining the limits of a population mean based on one sample for any level of confidence is presented as equation 6.04.

$$\mu = \bar{X} \pm Z(SE_M), \tag{6.04}$$

where Z is a Z score that will produce the desired probability of error (i.e., $Z = 1.65$ for $p = .10$, 1.96 for $p = .05$, and 2.58 for $p = .01$).

Table 6.1 Corresponding Values for Z, LOC, and *p*

Z	LOC	p
1.0	68%	.32
1.65	90%	.10
1.96	95%	.05
2.58	99%	.01

LOC = level of confidence.
p = probability of error for a two tailed test (sum of both tails of the curve).

An Example Using Statistical Inference

A researcher was interested in the average height of first-grade children in a school district. Eighty-three students randomly selected from throughout the district were measured with the following results: $\bar{X} = 125$ centimeters and $SD = 10$ centimeters. The population height was estimated at the 95% level of confidence, $p = .05$, as follows:

$$SE_M = \frac{10}{\sqrt{83}} = 1.1$$

and

$$\mu = 125 \pm 1.96\,(1.1) = 125 \pm 2.2, \quad p = .05$$

The researcher concluded with 95% confidence that the mean height of all the first-grade children in the school district was between 122.8 and 127.2 centimeters ($122.8 \leq \mu \leq 127.2$). There is, however, a 5% chance ($p = .05$) that this conclusion is incorrect.

In kinesiology, as in most behavioral sciences, the most common minimum level of confidence used is 95% ($p = .05$). But there is a developing trend to accept some research at the 90% ($p = .10$) level. The researcher decides which level to use, but the reader of the research must be the ultimate judge of what is acceptable. By consulting table A.1 in appendix A, we can determine any level of confidence and its equivalent Z. The decision of which level to use is based on the consequences of being wrong.

In medical research, if an incorrect conclusion may result in serious injury or death to the patient, then a very high level of confidence is desired. Even the 99% ($p = .01$) level may not be sufficient. The reader may wonder, Why not always use $p = .01$, since it is the least likely to result in error? Because, while the prediction of the population mean is more accurate at $p = .01$ than $p = .05$, it is more broad, thus less precise. In statistics, if you want less error, you must sacrifice precision. When an incorrect conclusion will not result in bodily harm or excessive financial loss, lower levels may be used. The user of the research must determine the consequences of being wrong in each case and accept or reject the conclusions accordingly. Franks and Huck (1986) provide an excellent review of the history and procedures for selecting a level of confidence. Chapter 8 discusses the pros and cons of selecting levels of confidence that are too low or too high.

Calculating Skewness and Kurtosis

A major assumption of the previous discussion about statistical inference is that the characteristics of the normal curve can be applied. Consequently, it is critical that we know if the data deviate from normality. **Skewness** is a measure of the bilateral symmetry of the data, and **kurtosis** is a measure of the relative peakedness of the curve of the data.

By observing the graph of the data and identifying the three measures of central tendency, we can get a general idea of the skewness of the data, but this method is not exact (see figure 4.1). Using Z scores, we can obtain a numerical value that indicates the amount of skewness or kurtosis in any set of data.

Because Z scores are a standardized measure of the deviation of each raw score from the mean, we can use Z scores to determine if the raw scores are equally distributed around the mean. When the data are completely normal, or bilaterally symmetrical, the sum of the Z scores above the mean is equal but opposite in sign to the sum of the Z scores below the mean. The positive and negative values cancel each other out, and the grand sum of the Z scores is zero.

If we take the third moment (the cube of the Z scores, or Z^3), we can accentuate the extreme values of Z, but the signs of the Z values remain the same. This places greater weight on the extreme scores and permits a numeric evaluation of the amount of skewness. Computing the average of the Z^3 scores produces a raw score value for skewness. The formula for calculating the raw value for skewness is

$$\text{Skewness} = \frac{\Sigma Z^3}{N}. \tag{6.05}$$

When the Z^3 mean is zero, the data are normal. When the Z^3 mean is positive, the data are skewed positive, and when the Z^3 mean is negative, the data are skewed negative. This effect can be seen by examining the data presented in table 6.2. Notice that the data are skewed negative. When these data are graphed (see figure 6.6), the skewness is easily observed.

Table 6.2 Calculation of Skewness and Kurtosis

X	Z score	Z^3	Z^4
5	1.08	1.26	1.36
5	1.08	1.26	1.36
5	1.08	1.26	1.36
4	0.27	0.02	0.01
4	0.27	0.02	0.01
4	0.27	0.02	0.01
4	0.27	0.02	0.01
4	0.27	0.02	0.01
3	–0.54	–0.16	0.09
3	–0.54	–0.16	0.09
2	–1.35	–2.46	3.32
1	–2.17	–10.22	22.17
		$\Sigma Z^3 = -9.21$	$\Sigma Z^4 = 29.80$

Mean = 3.67
SD = 1.23
N = 12

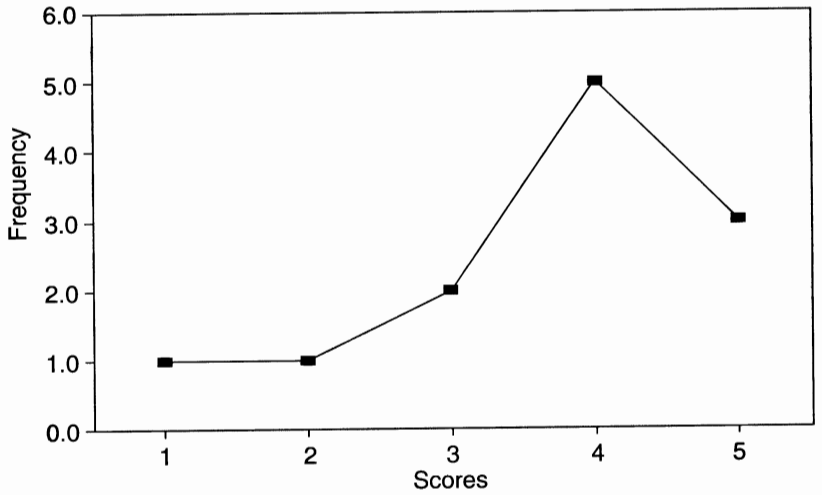

Figure 6.6 Negative skew.

Kurtosis may also be calculated from Z scores. By taking the fourth moment (Z^4) of the Z scores, the extreme Z values are again accentuated, but the signs are all converted to positive. When the average of the Z^4 value is 3.0, the curve is normal. To make the units equal for both skewness and kurtosis, the mean of Z^4 is typically reduced by 3.0. The formula for calculating the raw value for kurtosis is (AndersonBell, 1989, p. 173; Spiegel, 1961, p. 91)

$$\text{Kurtosis} = \left(\frac{\Sigma Z^4}{N} \right) - 3.0. \tag{6.06}$$

A score of 0 indicates complete normal kurtosis, or a mesokurtic curve, just as a score of 0 for skewness indicates complete bilateral symmetry. When the raw score for kurtosis is greater than 0.0, the curve is leptokurtic (more peaked than normal), and when the raw score is less than 0.0, the curve is platykurtic (more flat than normal).

Raw skewness and kurtosis scores are not easily interpreted, because a raw score alone does not indicate a position on a known scale. But when raw scores are converted to Z scores, they are easy to interpret. To convert the raw scores for skewness (equation 6.05) or kurtosis (equation 6.06) to Z scores for skewness or kurtosis, we divide the raw scores by the standard error. According to Dixon (1990, p. 137) the standard error for skewness is

$$SE_{\text{skew}} = \sqrt{\frac{6}{N}},$$

and the standard error for kurtosis is

$$SE_{\text{kurt}} = \sqrt{\frac{24}{N}}.$$

If we divide the raw scores for skewness or kurtosis by the appropriate standard error, we obtain a Z_{skew} or Z_{kurt} value:

$$Z_{\text{skew}} = \frac{\Sigma Z^3 / N}{\sqrt{6/N}}, \tag{6.07}$$

$$Z_{\text{kurt}} = \frac{\Sigma Z^4 / N - 3.0}{\sqrt{24/N}}. \tag{6.08}$$

These values may be interpreted as Z scores (i.e., values greater than 1.96 or less than –1.96 exceed $p = .05$, and values greater than 2.58 or less than –2.58 exceed $p = .01$). Typically, data are considered to be within acceptable limits of skewness or kurtosis if the Z values do not exceed ± 2.0.

Using the data from table 6.2, we can find Z_{skew} in the following manner:

$$\text{Skewness} = \frac{-9.21}{12} = -0.77$$

$$SE_{\text{skew}} = \sqrt{\frac{6}{12}} = 0.71$$

$$Z_{\text{skew}} = \frac{-0.77}{0.71} = -1.08$$

Z_{kurt} can be found as follows:

$$\text{Kurtosis} = \frac{29.80}{12} - 3.0 = -0.52$$

$$SE_{\text{kurt}} = \sqrt{\frac{24}{12}} = 1.41$$

$$Z_{\text{kurt}} = \frac{-0.52}{1.41} = -0.37$$

From these values ($Z_{skew} = -1.08$ and $Z_{kurt} = -0.37$) we can determine that the data in table 6.2 and figure 6.6 are slightly skewed negative and slightly platykurtic; however, neither value approaches significance (±2.0). So we may conclude that the data are within acceptable ranges of normality. Note that data sets with small values of *N* may appear to be significantly skewed when graphed (see figure 6.6), but the true evaluation of the degree of skewness must by made by *Z* score analysis.

Summary

Raw scores may be converted to standard scores in the form of *Z*, percent, *T*, or stanine to provide more information about the data and to assist in evaluating raw data. Standard scores are also useful for comparing the results of tests measured on different units of measurement (e.g., comparing time to force or distance). Because standard scores have a common central tendency and variability, data presented in standard score units may be directly compared regardless of the unit of measurement of the raw score.

By using the characteristics of the normal curve, estimates of the parameters of populations may be made from sample statistics. Calculations can also be performed to determine the standard error of the mean, which indicates the amount of error in an estimate of a population mean based on one random sample.

The assumption underlying the process of statistical inference is that the characteristics of the normal curve can be applied to the data. We can determine how much the distribution of a data set deviates from normality by calculating both skewness and kurtosis.

Problems to Solve

1. Find the percentage of values that falls between the mean of a given set of population data and a *Z* score of +0.35. (Hint: Use appendix A, table A.1).
2. Given a set of data with a mean of 25.7 and a standard deviation of 5.2, calculate the *Z* score equivalents of the following raw scores: (a) 21.6, (b) 28.9, and (c) 24.5. What is the interpretation of the *Z* scores?
3. Use table A.1 in appendix A to determine the equivalent percentile rank of each of the scores in problem 2.
4. Convert the *Z* scores calculated in problem 2 to *T* scores.
5. What is the stanine value for each of the scores in problem 2? (Hint: Use figure 6.5).

6. Using the following sample data, compute the standard error of the mean. What can you say about the value of SE_M in relation to the size of the *SD* and *N*?

	SD	*N*
a.	21.4	36
b.	2.1	106
c.	56.7	19

7. Suppose you wanted to know the mean $\dot{V}O_2$max in milliliters per kilogram per minute of all females at a university. You randomly selected 50 names from the files of the administration office and invited the subjects to be tested in the lab. Thirty-five accepted the invitation and were tested. The mean $\dot{V}O_2$max of this sample was calculated to be 38.5 with a standard deviation of 4.7. If you wanted the population mean estimate to be accurate at the 95% level of confidence, what would be the upper and lower limits of the predicted $\dot{V}O_2$max value?
8. Using the best random sample you can get, collect data and estimate the average height in inches or meters of all members of your class. You select the appropriate level of confidence. Look at your hand calculated values (especially the maximum, minimum, mean, and standard deviation) to see if they make sense. It is always a good idea to "eyeball" your answers. If time permits, measure the total population (all members of the class). How close did your estimate come to the true population mean?
9. Enter the data from the sample you collected in problem #8 into a computer statistical package. Instruct the computer to calculate: mean, maximum, minimum, range, standard deviation, standard error of the mean, *Z* scores for each value, skewness, and kurtosis. Using the values generated by the computer, estimate the population mean (μ). How does the computer output compare with your hand calculated values in problem #8?

See appendix C for answers to problems.

Key Words

Alpha
Central limit theorem
Inferential statistics
Kurtosis
Level of confidence
Probability of error
Sampling error
Skewness
Standard error of the mean
Stanine
T score
Z score