Chapter Three

Data Management in Measurement

In measuring achievement or proficiency in the use of language, we usually administer one or more tests to one or more students. What we can infer about student competence or ability depends on (1) what we can discover about the characteristics of the testing instrument(s) and (2) what we can interpret as the significance of performance scores on the instrument(s). These two important steps in measurement require facility in the management of scores of the test and its component items. Accordingly, this chapter is devoted to the management of test data.

3.1 Scoring

Tests with Items

Most language tests are composites of individual items, and scores on these tests are obtained by summing the correct answers. Each item response is scored as either correct or incorrect. Typical varieties of these tests include *multiple-choice*, *true/false*, *cloze completion*, or *question-answer* types of items. Because it is possible to guess the correct answers with multiple-choice or true/false type items, correction-for-guessing procedure is often used. The net result is that measurement error due to guessing is reduced, and test scores become slightly more reliable. One common formula applied in correction-for-guessing procedure is the following:

$$S_{cg} = N_r - \frac{Nwa}{N_o - 1} \qquad (3.1)$$

where, S_{cg} = the score corrected for guessing
N_r = the number of right item responses

N_{wa} = the number of wrong items attempted (Note that this does not count omitted items.)

N_o = the number of options available per item

As is obvious from the formula, this procedure introduces a weighted penalty for responding without comparative certainty. Typically the amount of reliability gained by this procedure is only substantial when items use few (i.e., two or three) options. With increase in response options, the advantages of correction for guessing diminish rapidly. Part of the reliability, or person-separability, of some tests is a function of speededness. In a test of reading comprehension, less capable students are partially distinguished from more capable students in that the less capable students are unable to finish the test. When correction for guessing is used for these tests, the contribution to reliability from speededness is partially removed. Slower, but more accurate, students are not penalized as much for not finishing the test when correction for guessing is used.

Latent trait approaches to scoring, as employed on the Test of English as a Foreign Language (TOEFL), and an increasing number of modern examinations have the potential of identifying and controlling for guessing. Scoring according to latent trait methodology does not require the sum of a person's correct responses but can be determined with reference to performance on a very small subset of items, the difficulties of which closely match the person's ability. Latent trait methodology is introduced in chapter eight.

Tests with Prose Passages

The scoring of prose passages produced by the students may introduce measurement error due to rater subjectivity unless appropriate precautions are taken. *Precis*, *composition*, and *dictation* tasks are common examples of this category of tests.

Dictation passages may be scored by treating each word as a separate item, counting it correct if present and incorrect if absent. However, research on this technique found improved reliability of scoring to be at the expense of validity (Shereaf, 1981). The more common procedure is to allot a maximum score possible, often corresponding to the number of words in the passage, and then systematically to subtract points for errors of grammar, spelling, or punctuation, depending on the purpose of the test.

Oller (1979) suggests two basic scoring techniques: *correct words-in-sequence* and *error counting*. By the latter technique, a student may receive a negative score if error penalties exceed the total mark possible. Shereaf (1981) provides a thorough discussion of statistical advantages and disadvantages of a variety of techniques of scoring dictation.

Similarly, free-writing tasks may be scored by deduction for errors from a maximum permissible score. However, the problems in scoring free-writing tests are manifold. Some examinees write longer passages than others and so produce more errors as a function of their greater effort. This suggests that rate of errors per specified passage length is more useful than an actual frequency tally of errors. Some examinees avoid all complex structures and sophisticated vocabulary for fear of mistakes, while others attempt more creative use of language and thus generate comparatively more errors. Such problems suggest that an element of subjective judgment on the part of the person scoring the test may be necessary. Subjectivity of this judgment may be minimized in at least four ways. *First*, a rating schedule may be used which operationally distinguishes between superior and inferior performance. One example is provided in Table 3.1.

Notice that this particular rating schedule permits a maximum of ten points for any given composition or essay. The schedule calls for equal weighting of mechanics and content, with one point awarded for satisfactory performance in each component area. Depending on the level of the particular student sample, the operational definition of acceptable or satisfactory performance will vary. One might say for a given sample length that more than two errors of spelling on the part of the examinee would result in forfeiting the possible one point awarded for spelling, etc. Using such a rating schedule tends to "objectify" the rater's task in the sense that ratings by various persons at various times will tend to reflect the same underlying criteria and thus become more consistent.

There is still a problem with such a scale; that is, that the selection of component areas and weights is somewhat arbitrary. Thus, one teacher may prefer to allow comparatively more weight for grammar usage, or for organization, and so on. A more scientific procedure for assigning appropriate weights to component parts of an examination or rating schedule is discussed under the multiple regression section of chapter ten. Another variety of rating schedule might specify in advance the kind of writing behavior that warrants specific marks. An example of such a schedule is provided in Table 3.2.

Here it may be desirable to allow partial scores, as 3.5, 2.5, etc., for persons whose performance is located between points on the scale.

A *second* way to reduce the subjectivity of scoring beyond the use of rating schedules is to insist on rater competence or expertise. This is to say that selection criteria be imposed in the choice of raters. Obviously no one should undertake the task of rating who does not know the language well or who has not had sufficient prior training and/or experience in rating.

A *third* possible procedure for reducing the subjectivity in the scoring of compositions or other written passages is to employ multiple independent raters. In this way every composition is read and marked independently by more than one

TABLE 3.1 Sample rating schedule* for use with precis, essay, and composition

MECHANICS		CONTENT	
Area	Weight	Area	Weight
Spelling	1	Organization	1
Grammar usage	1	Relevance to topic	1
Punctuation	1	Creativity/interest	1
Orthography	1	Range and sophistication of syntax	1
Paragraphing	1	Richness of vocabulary/expression	1
	5		5

*Other more sophisticated schedules are available (Cf. Jacobs et al.).

TABLE 3.2 A sample behavior-specific rating schedule for use in foreign language writing evaluation

BEHAVIOR	RATING
1. Writing is indistinguishable from that of a competent native speaker.	5
2. Writing is grammatically correct but employs nonnative usages.	4
3. Writing contains infrequent errors of grammar, lexis, spelling, or punctuation.	3
4. Writing contains numerous errors of grammar, lexis, spelling, and punctuation.	2
5. Writing is incomprehensible. Orthography is illegible.	1

rater. Such ratings must be independent in the sense that raters do not confer nor see the mark assigned by the other rater(s) before they have made their own judgment. The ratings of each rater for each examinee are then totaled or averaged to give a composite score for each examinee that reflects the combined judgments of all raters.

It is often desirable to employ all three of these "objectifying" procedures simultaneously. Thus, an ideal procedure is to use a rating schedule with multiple independent raters who are experts. Mansour (1978) found that the use of multiple independent raters who did not employ rating schedules necessitated the use of as many as seven raters rating each composition in order to attain acceptable reliability of scoring.

A *fourth* "objectifying" procedure is to elicit multiple writing samples from the examinee, preferably at various times on various topics. This would control for the fact that writing ability may vary with topic and time of day, etc.

Tests of Oral Communication

Many comments already made on scoring written passages apply equally well to scoring oral communication or production performance. It is equally desirable to have rating schedules and multiple independent raters who are experts. It is equally preferable for raters to judge performance in more than one topic or situational context. In addition to these scoring procedures, it may be desirable to record examinee performance on cassette or video-cassette in order to permit repeated or more thorough judgment by the rater(s) involved. Use of recording equipment must be considered thoroughly beforehand, not only to ensure quality of recording but also to determine whether or not subjects find such equipment threatening or intimidating and whether resultant speech samples become unnatural or otherwise invalid.

A sample rating schedule for use in judging speaking ability in a foreign language is reported in Table 3.3. In it the interviewer or rater records the name of the examinee and then proceeds to rate examinee performance on a scale of 1 to 5, depending on how nearly it approximates native-speaker performance. Further information on techniques of scoring samples of oral production, such as interviews, is available from Foreign Service Institute and Educational Testing

TABLE 3.3 A sample rating schedule for rating oral communication performance

Name of Examinee	Fluency (1–5)	Pronunciation Accuracy (1–5)	Grammar Accuracy (1–5)	Expressive Content (1–5)	Total (4–20)

Service (1962). The present grouping of tests with prose passages and tests of oral communication is not intended to imply that writing and speaking may not be tested well using tests with items rather than tests with extended free samples of language. The classifications mentioned have been introduced because of their usefulness in promoting understanding of scoring procedure. Additional rating schedules have been presented by Harris (1969) and others.

3.2 Coding

Once a set of dependable scores is in hand, reflecting the performance on a test or a battery of tests for a sample of examinees, the task becomes one of organizing and recording this raw test data. For tests with items, it is helpful to record data on a *scoring matrix*, normally on computer coding sheets or ordinary graph paper. An example is provided in Table 3.4.

Notice that the numbers on the left correspond to the identification numbers assigned to the papers of the students (p). The numbers across the top correspond to the 15 individual items (i) on the vocabulary subtest of the Ain Shams University Proficiency Examination (ASUPE) (Henning, 1978). Actual item performance data is recorded in the matrix with ones and zeroes. A *one* signifies that the examinee got that particular item correct, while a *zero* means that the examinee response was incorrect.

The next step in working with such a scoring matrix is to obtain the marginal totals. The totals at the right represent the total subtest scores for each student. The totals at the bottom of the matrix signify the number of persons getting correct scores on each item. As such they provide one measure of the comparative difficulty of the items.

Use of such a matrix is essential in the analysis of test data, particularly when there are large numbers of items and examinees and reliance must be placed upon a computer. Even when item data is not being considered, it is useful to resort to scoring matrices for recording subtest total scores for each examinee.

TABLE 3.4 A scoring matrix based on ASUPE Form A vocabulary results for third-year university students

		ITEMS (i)															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	TOTAL
THIRD YEAR STUDENTS (p)	1	1	1	0	1	0	0	1	0	1	1	1	1	0	1	0	9
	2	1	1	1	1	0	0	0	0	1	1	0	1	0	0	0	7
	3	1	1	1	0	1	1	1	1	1	0	1	0	0	0	0	9
	4	1	0	1	1	1	1	1	1	1	1	1	1	0	0	0	11
	5	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	14
	6	0	0	1	0	1	0	0	0	0	1	1	1	0	0	0	5
	7	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0	12
	8	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	13
	9	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1	12
	10	0	0	0	0	1	0	1	1	0	1	1	0	0	0	0	5
	11	1	0	0	0	1	0	0	0	1	1	1	1	0	1	0	7
	12	1	1	0	1	0	0	0	0	1	1	1	1	1	0	0	8
	13	1	0	0	0	1	1	1	1	0	1	1	1	0	0	1	9
	14	1	1	1	0	1	0	0	1	0	0	0	1	0	1	0	7
	15	0	1	0	0	0	1	0	0	1	0	1	1	0	1	0	6
	16	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0	9
	17	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	3
	18	0	1	1	1	1	1	0	0	0	0	1	0	0	0	0	6
	19	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0	12
	20	1	1	0	1	1	1	0	0	0	0	0	1	0	0	0	6
	21	1	1	0	0	1	0	0	1	1	0	0	0	0	0	0	5
	22	0	1	1	1	1	0	1	1	1	0	0	0	0	0	0	7
	23	1	0	0	0	0	0	0	0	0	0	1	1	1	1	1	6
	24	0	1	1	0	1	0	1	0	1	0	1	1	0	1	0	8
	25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	14
	26	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	14
	27	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	3
	28	0	1	1	0	1	1	1	1	1	1	0	0	0	0	0	8
	29	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	4
	30	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	9
TOTAL		20	23	18	18	23	15	14	17	19	18	19	19	8	13	4	

3.3 Arranging and Grouping of Test Data

Range and Distribution

In order to better understand the nature of the test and the comparative abilities of the examinees, it is desirable to arrange the scores in such a way that we can readily visualize the overall results. If we arrange the total test scores from Table 3.4 in this way, the result would appear as in Table 3.5.

TABLE 3.5 A frequency distribution of scores from the ASUPE Form A Vocabulary Subtest

Score	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Frequency	0	0	0	2	1	3	4	4	3	5	0	1	3	1	3	0

Notice from Table 3.5 that no students obtained a score of less than 3 or more than 14. We could say therefore that the *actual range* of scores was 3–14, or 11. The *possible range* was 0–15, or 15. Since this was a multiple-choice type test with four options, one would expect the examinees to get a score of at least 25 percent by mere guessing. Thus, a score of 3.75 or below would actually be meaningless for discriminating among students in the ability being tested. Therefore we conclude that the *effective range* of this test was 4–14, or 10. Generally tests with the broadest possible effective range are most useful in discriminating among examinees on the ability under consideration.

The information presented in Table 3.5 is called a *frequency distribution* because it reports the frequency with which examinees obtained particular scores. This same information might have been presented graphically as in Figure 3.1.

Several kinds of frequency distribution are of interest in the field of test development or measurement theory. One kind of scoring distribution is called *skewed* distribution. Examples of this type of distribution are provided in Figure 3.2.

Notice from the example of *positive skew* at the top of Figure 3.2 that, in this case, the examinees found the examination too difficult. Most of the student scores are clustered near the zero point of the scoring range. In the case of the distribution with *negative skew*, just the opposite is true. Here the examinees found the test too easy, and a high percentage of them obtained a perfect score. Negative skew is not always considered a problem. In the case of criterion-referenced tests, this type of distribution is desirable as an indication that a majority of the students attained the objective of the program of instruction.

Another useful concept in the consideration of the shape of a scoring distribution is the notion of *kurtosis*, or peakedness. Figure 3.3 illustrates two extremes of kurtosis called a *leptokurtic* or peaked distribution and a *platykurtic* or flat distribution.

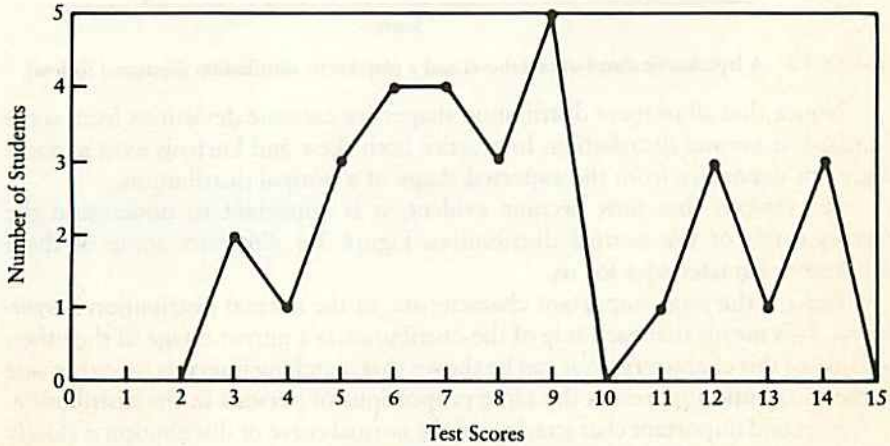

FIGURE 3.1 A frequency distribution of ASUPE scores

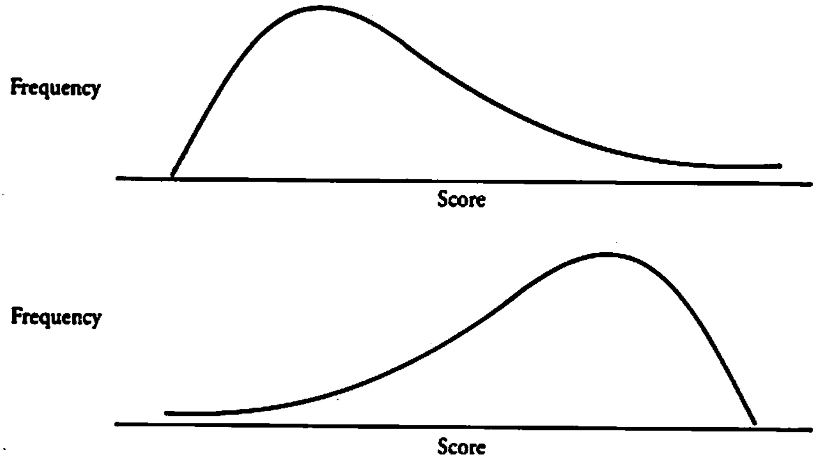

FIGURE 3.2 Typical shapes of skewed distributions. Positive skew is illustrated above, and negative skew is shown below.

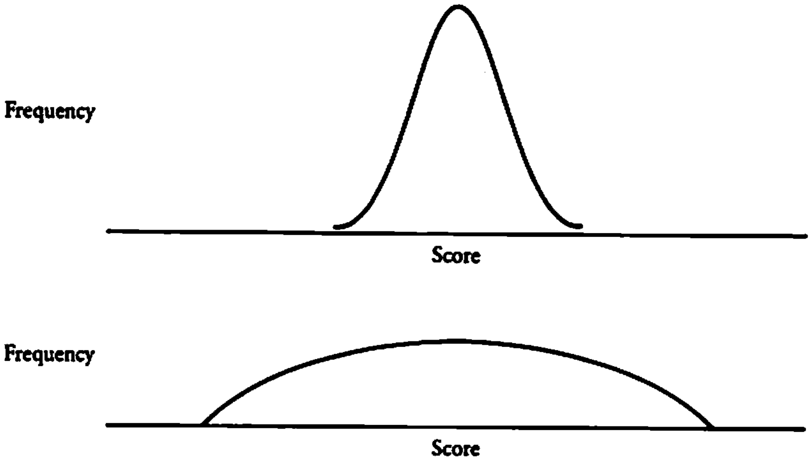

FIGURE 3.3 A leptokurtic distribution (above) and a platykurtic distribution illustrated (below).

Notice that all of these distribution shapes are extreme deviations from some standard or *normal distribution.* In practice both skew and kurtosis exist as some degree of departure from the expected shape of a normal distribution.

For reasons that now become evident, it is important to understand the characteristics of the normal distribution. Figure 3.4 illustrates some of these important characteristics for us.

Perhaps the most important characteristic of the normal distribution is *symmetry.* This means that each side of the distribution is a mirror image of the other. Because of this characteristic, it can be shown that matching intervals on either side of the distribution represent the same proportions of persons in the distribution.

A second important characteristic of the normal curve or distribution is closely related to the first: that, for the normal distribution, the *mean*, the *median*, and the

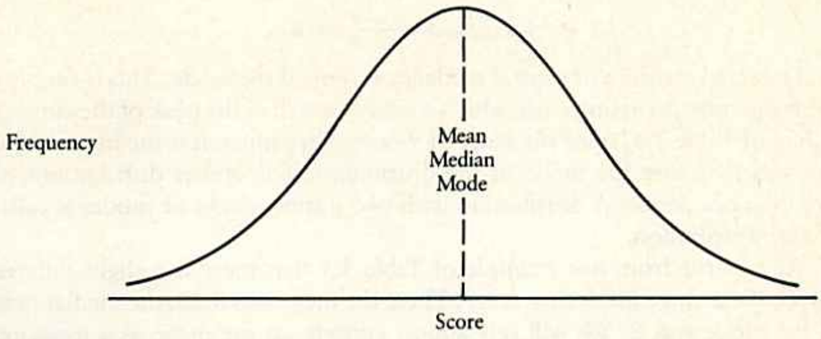

FIGURE 3.4 The normal distribution

mode fall at the same point on the scoring axis. These three measures of central tendency or numerical average are discussed in greater detail in the following section.

The third property of the normal distribution is that the curve is *asymptotic*; that is, it never intersects the horizontal axis. This means that in a theoretical sense there is always the possibility that some one person's score may extend infinitely far to the right or to the left on the horizontal axis. In practice, however, because we cannot test an infinitely large sample and because tests have a finite number of items, this characteristic is not so important to us as the other two properties of the normal distribution.

Measures of Central Tendency or Statistical Average

There are at least three ways to locate a mid-point of the distribution. The most common method is by using the arithmetic *mean*. To obtain the mean we sum the individual scores of the distribution and divide by the total number of scores in the distribution.

This relationship is shown in the following formula:

$$\bar{X} = \frac{X}{N} \qquad (3.2)$$

where, $\bar{X}$ = the arithmetic mean
X = the sum of all scores
N = the number of persons in the sample

Applying this formula to the data of Table 3.5, we obtain the following:

$$\bar{X} = \frac{248}{30} = 8.27$$

A second measure of central tendency is termed the *median*. This is the numerical point in the distribution at which half of the obtained scores lie above and half below. With an odd number of scores in the distribution, this is simply the centermost score. When there is an even-numbered distribution, as in Table 3.5, the median is the midpoint of the interval between the two centermost scores. This can be determined by summing the two centermost scores and dividing the result by two, as in the following example: from the data of Table 3.5

$$\text{Median} = \frac{8+8}{2} = 8$$

The third measure of central tendency is termed the *mode.* This is simply the most frequently occurring score which is easily located as the peak of the curve. For the data of Table 3.5, since the score of 9 occurs five times, it is the most frequent score and therefore the mode of the distribution. Sometimes distributions have more than one mode. A distribution with two distinct peaks or modes is called a *bimodal* distribution.

We observe from our example of Table 3.5 that there is a slight difference between these three kinds of average. There the mean was 8.23, the median was 8, and the mode was 9. We will rely almost entirely on the mean as a measure of central tendency in this text.

Measures of Dispersion or Variability

Another concern in describing the shape of a distribution is to quantify the density or *dispersion* of the scores around the mean. Consider the leptokurtic distribution at the top of Figure 3.3. Most of the scores are closely gathered around the mean. Such a distribution is said to have low *variability.* By contrast, the platykurtic distribution at the bottom of the same figure is said to have high variability because the scores are spread widely.

Two measures of variability are commonly used in test development. The first is called the *variance.* It is determined for a sample of scores by the following formula:

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{N-1} \tag{3.3}$$

where, s^2 = the variance
X = any observed score in the sample
$\bar{X}$ = the mean of all scores
N = the number of scores in the sample
Σ = the symbol for summation

The second measure of variability is called the *standard deviation.* It is obtained by taking the square root of the variance, as shown by the following formula:

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}} \tag{3.4}$$

It is important to note that these formulas are designed to provide an unbiased estimate of the variability in an underlying *population* of scores based on a *sample* of scores drawn from that population. If it were possible for us to obtain or conceptualize all of the scores in the population, we would rely on slightly different formulas: i.e.,

$$\sigma^2 = \frac{\Sigma(X-\bar{X})^2}{N} \tag{3.5}$$

$$\text{and } \sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} \tag{3.6}$$

Notice that for *population parameters*, as opposed to *sample statistics*, reliance is placed upon Greek rather than Latin symbols and N replaces $N-1$ in the denominator.

3.4 Summary

In this chapter we considered management of test data, including a review of scoring procedures for a variety of language tests. Coding test data through the use of scoring matrices has been discussed. And the grouping of test data according to the range and shape of the distribution of scores was the final general topic. Skew and kurtosis have been considered along with the properties of the normal distribution. Three measures of central tendency and two measures of variability or dispersion were introduced. A final comment was offered on the difference between population parameters and sample statistics.

Exercises

1. What would the score be after correction for guessing for a student who got 72 out of a possible 90 correct on a three-option multiple-choice test? Assume the student attempted all but five of the 90 items.
2. What are three techniques which may be used to reduce subjectivity of scoring for compositions or oral interviews?
3. What are some advantages of a scoring matrix?
4. How does the *effective range* differ from the *possible range* or the *actual range*?
5. What does the presence of positive and negative skew in the distribution of scores tell us about the nature of the test?
6. Given the following set of scores from a 25-item reading comprehension test: 11, 7, 13, 9, 14, 6, 17, 12, 14, 14, 10, 13, 12, 10, 19, compute the mean, the median, and the mode.
7. Name three properties of a normal distribution.
8. Using the data from problem 6, compute the standard deviation and the variance.
9. What is the difference between a population parameter and a sample statistic?
10. In a normal distribution, what proportion of the scores lie between plus and minus one standard deviation? Between plus and minus two standard deviations?