\*\*Note: Remember to use Musial as reference.

Musial, D., Nieminen, G., Thomas, J., & Burke, K. (2008). Foundations of meaningful educational assessment. Boston: McGraw Hill.

CHAPTER 11 Statistical Applications to Assessment

Chapter Objectives

After reading and thinking about this chapter, you will be able to:

• Describe and calculate statistical concepts commonly used by classroom teachers, such as central tendencies and measures of variability.

• Explain the special characteristics of the normal distribution.

• Describe typical item analysis procedures and how they can be used to ­inform assessment-based decisions.

If you score a 95 on a math test, would you be pleased? Your first response might be, “Of course!” But what if there were 200 possible points? Or, if it were a 100-point test, would you be pleased to know that a 95 was the lowest score in the class? The score on an assessment tells only part of the story. To be meaningful, the score must be interpreted with respect to other variables, such as the scores of other students, the student's prior performance on similar assessments, the content of items answered correctly and incorrectly, and so on. In this chapter, we will introduce you to statistical concepts that will help you interpret scores in a meaningful way; but, in doing so, we urge you to think about how your interpretation might differ from the interpretation of another teacher, the student, or the student's parents.

This chapter introduces you to basic statistical concepts that can help you describe student performance on assessments and help you evaluate the effectiveness of your instruction. Students are often most interested in their own performance, but their teachers need to know more. Teachers can gain insight about the effectiveness of their instructional approaches based on the way the entire class performs on an assessment. This chapter will give you an understanding of several common statistical procedures that 299300can be used to look for patterns across the whole class of students. It will provide the tools to compare one student's performance to the performance of the whole class, to determine if your learning targets have been met, and to evaluate the assessment itself.

Statistics provide pictures or insights about groups and provide information about individuals. Examining patterns across many students is very helpful for determining the effectiveness of an instructional method. If an instructional approach helps a large number of students, it is worth continuing. Of course, those students who do not benefit from the approach need to be provided other options. Foundational concerns about using statistics stem from the delicate balance between the importance of the individual versus the importance of the group. As teachers we must always be wary of making decisions that only benefit the larger group at the sacrifice of the individual student.

Foundational Questions for Your Consideration

• When is it appropriate for a teacher to compare a student's performance on an assessment to the performance of other students?

• When is it useful to know how a class performed overall? Would that knowledge provide insight into how well students learned and how effective instruction was?

• How do you know if an assessment item achieved its purpose?

Looking Beyond the Test Score

As we begin our discussion of statistical applications to classroom assessment, let's take a look at a typical classroom. The community the students come from is diverse. About half of the students are male and half female, with varying socioeconomic backgrounds. The school racial/ethnic makeup is 25.5 percent White, 35.2 percent Black, 38.4 percent Latino, 0.8 percent Asian, and 0.1 percent Native American. Fifty-three percent are low income, and 18.9 percent are English language learners. There are 26 students in the classroom.

The class has just completed an American history unit on the Great Depression, culminating in a summative assessment, a 50-item multiple-choice unit test. Each item is worth 1 point, so the highest possible score is 50 points. The teacher, Ms. Dailey, scores each student's test and enters the score in the grade book as shown in Figure 11.1.

Figure 11.1 Student Scores on the Great Depression Unit Test

300 301

Ms. Dailey is aware that some students did not perform as well as she expected or hoped. She notices this when hand-scoring the test and when entering the scores into the grade book. She also notes a few students did exceptionally well. She is not sure what to make of the assessment information. Did all students meet the learning objectives? Is there a pattern to the assessment data? That is hard to tell based on the list of scores in the grade book. Should she reteach part of the unit or move on to the next unit? Should she provide additional instruction for some students, but not all? Ms. Dailey is having a hard time deciding what to do. This is a common situation for many teachers.

An additional factor is weighing on Ms. Dailey's mind. While scoring the test, she noticed that some of the items were answered correctly by all or almost all the students, a good thing from her perspective. She is also aware that a number of students performed poorly on specific items. She is concerned that perhaps the class did not learn certain aspects of 301302the unit. Again, Ms. Dailey is not sure what to make of her observations and the data.

There are often subtle patterns of correct and incorrect answers on an assessment. It is difficult to detect such patterns when scoring and grading assessments, especially with a large class. If Ms. Dailey were able to detect these patterns, she might be able to place individual and whole-class performance in a better context. That level of detail and analysis could be useful in making good instructional decisions.

Fortunately, there are tools available to the teacher to help determine what typical student performance is for the class based on their distribution of test scores, how spread out these scores are, and how an individual fits into that spread. There are also tools to analyze the quality of test items that can tell the teacher how easy or difficult an item is and how well an item measures a concept or construct.

Statistical Concepts

Let's think about Ms. Dailey, the assessment data she has, and how she could use the data. She could use the data to

• See if students met learning objectives.

• Gauge the effectiveness of her instruction.

• Plan future instruction.

For example, it would be useful to know what the typical student performance was on the assessment. Overall, did students perform well or poorly? How spread out are the scores? What is the lowest score? What is the highest?

Distribution of Scores

A good place to begin detecting patterns in scores is to consider what the data really look like. We need to see the data in some organized way to extract meaning from it. It would benefit Ms. Dailey if she were able to detect patterns in the distribution of scores.

Frequency Table

Many teachers begin looking at test scores and other numerical assessment data by arranging them in a frequency table. This lets us get a sense of the data, to look for trends. The easiest and quickest way to do this is to arrange the data in a list from lowest score to highest score. We could then tally the number of times a particular score occurred. Figure 11.2 is Ms. Dailey's unit test data arranged from lowest score to highest score. The tally marks indicate the frequency of each score. Note 302303the use of tally marks instead of Arabic numerals ( instead of 6). This helps us see the shape of the data. What shape do you see?

Ms. Dailey could begin interpreting her test scores using this frequency table. For example, she could see where a particular student performed compared to other students. She could more easily see what the highest scores are, what the lowest scores are, and what the middle scores tend to be. She could also see that a few students performed poorly on the assessment, while the majority of the students did very well.

303 304

Histogram

Another useful way to display a frequency distribution is to create a histogram, a pictorial representation of data in the form of a bar graph. Each score can be listed on the horizontal axis (the X-axis), and the tally or frequency of each score can be displayed on the vertical axis (the Y-axis). Figure 11.3 shows the histogram of the assessment scores earned by the students on the unit test. The histogram shows roughly the same shape as the tally marks in the frequency table in Figure 11.2. (Note how rotating the frequency table of tally marks in Figure 11.2 counterclockwise gives you the same view as the histogram.)

Figure 11.3 Histogram of Student Scores on Great Depression Unit Test

There is a lot more meaningful information conveyed in a histogram than in an unorganized set of assessment scores. We can already see several patterns from this pictorial representation. Again, it is easy to note how a few students performed poorly on the assessment in contrast to the majority of the students who earned scores of 42 or more. Do you see the really low scores of 22, 23, and 27? Also, it looks as if most students did fairly well, given the cluster of students scoring 42 or better on the 50-item test.

Besides looking at individual scores, Ms. Dailey could group scores together into intervals that might represent specific grades earned by students. Ms. Dailey uses the grading scale shown in Figure 11.4 to assign letter grades to unit test scores. With the grading scale shown in the figure, a student earning a score of 48 would earn an A, a student earning a 42 would earn a B, and so on.

We could display the frequency of each grade in a histogram by counting how many students received a particular grade on the unit test. Figure 11.5 shows what the unit test grades would look like.

Figure 11.5 Histogram of Student Grades on Great Depression Unit Test

304 305

Histograms provide a way of visually detecting patterns in assessment data. A quick glance at the histogram of the grades suggests a number of students in Ms. Dailey's class performed quite well. More than half the students earned either an A or B. Of course, this is the same pattern we saw in the histogram of individual scores in Figure 11.3.

Frequency Polygon

Another useful representation of assessment data is the frequency polygon, a line graph similar to the histogram. Test scores are again written along the X-axis, while the frequency of each score is on the Y-axis. A dot is placed at the intercept of the midpoint of the interval for a particular score (above the X-axis) and the frequency of that particular score (along the Y-axis). A line is drawn between adjacent dots. A frequency 305306polygon created from the Great Depression unit test scores in Ms. Dailey's class is shown in Figure 11.6.

The shape of the frequency polygon in Figure 11.6 is similar to the shape of the histogram in Figure 11.3. Frequency polygons and histograms essentially convey the same information. Ms. Dailey could use either to gain a better understanding of her assessment data.

Now that she has the frequency table, the histogram, and the frequency polygon, Ms. Dailey is in a position to begin to answer the question, “What is the typical student performance?” And, “How well did I do in helping students meet the learning outcomes?” How would you answer these questions? While it is possible to get a rough idea by observing the pictorial representations of the set of scores, the answer would lack precision. And that leads to another question, namely, “What exactly is typical performance?”

Measures of Central Tendency

There are several ways to measure typical performance numerically. The statistical term for typical performance is central tendency, which is a numeric summary of a set of scores. Think of central tendency as a measure of where data tend to cluster together. There are three common measures of central tendency: mean, median, and mode. Each of these is a different way to summarize the scores into a single number.

Mean

The mean is already familiar to you. It is simply the arithmetic average of a set of scores. As you recall, the average is calculated by 306307taking all the individual scores, adding them together, and dividing by the total number of scores. The formula is written as

which you read as, “The mean equals the sum of all scores divided by the number of scores.”

Σ is the capital Greek letter sigma and stands for the sum of a group of numbers. X represents each individual score and N is the total number of scores. So, what is the average (mean) score of Ms. Dailey's class on the unit test? Here is how the mean is calculated:

The mean score is 40.7, or approximately 41. Ms. Dailey could use this measure of central tendency to characterize the typical performance of her class. Overall, the typical student answered about 81 percent of the items correctly (40.7 items correct divided by the total of 50 is 0.81 or 81%). She could also get an idea of how individual students performed compared to the typical student. For example, it looks like Maria Stephano did very well compared to the average student in class. Maria's score of 48 is clearly above average. Robert Etner, with a score of 23, appears to have performed well below the average. Lucinda Fuentes and Belinda Williams appear to be typical, with scores of 40 and 41 respectively, right at the mean.

Note that the mean uses all scores in the set of data. Every assessment score is used to calculate the mean, including those who did extremely well and those who did extremely poorly. Take a look at the frequency table again in Figure 11.2. How did Robert Etner's score of 23 and Joseph Trapo's score of 22 impact the mean? And Wilma Bennett's score of 27 was also lower than the rest of the class. A look at the frequency table will show that most scores seem to be in the range of 37 to 50, while those below 33 certainly seem to be unusual. Scores that are quite different from the majority (either higher or lower) are called outliers. Could these outliers be distorting the mean by pulling it lower than what might be the typical or average performance on this test?

If these three students (Robert, Joseph, and Wilma) had not taken the test, their low scores would not have been calculated into the class mean. The average score, if we recalculated it, would be 43 instead of 40.7. This difference may seem small, but it points out that the mean is influenced by 307308all assessment scores, including outliers. Since all data points are included in the mean, we must be careful of our interpretation of the mean. It may or may not represent what is considered to be typical student performance. Always look at the set of scores once they are arranged in a frequency table or a histogram. Figure 11.7 illustrates this.

Only by examining the data in an organized way would we have a sense of how the scores below 33 are skewing or pulling the mean lower than we might expect. A skewed distribution that is pulled lower by outliers is a negatively skewed distribution. A distribution that is pulled higher by outliers is a positively skewed distribution. Did you get the clues? Positive distributions pull toward the more positive end, while negatively skewed distributions pull toward the more negative end. (In this usage, positive does not necessarily mean good, it simply means a higher number.)

Figure 11.8 shows three distributions; a positively skewed distribution, a normal distribution, and a negatively skewed distribution. Note how the tails pull the distribution in a particular direction. Knowing the shape of the distribution can have important classroom implications. For example, at the beginning of a unit of instruction, student performance on a pretest might take the shape of a positive distribution (most students perform poorly and have not mastered the instructional material). That is to be expected. At the end of instruction, it would be desirable that most students perform quite well on the end-of-unit assessment. In other words, the shape of the distribution becomes negatively skewed. This would indicate that most students learned the material and met desired expectations.

Seeing a change in the shape of a distribution from a positive skew to a negative skew would be a good thing in the classroom. A shift in skew suggests there has been a growth in student learning. Figure 11.9 308309represents a shift in student performance between a pretest at the beginning of a unit of instruction and a posttest at the end of instruction if teaching was effective and students met desired learning outcomes.

Figure 11.9 Expected Shift in Skewness When Teaching Is Effective and Students Have Met Desired Outcomes

Median

A second measure of central tendency is the median, the middle score in a set of scores. Half of the scores are above the median, and half are below the median. The median represents the score of the individual who would be right in the middle of the set of scores. To find the median, you would first arrange the scores from lowest to highest. Then determine 309310which score is in the middle, with half of the scores below it and half above. That score is the median. For an example, look at the following score set, which has been arranged from lowest score to highest score.

The middle score is 29. Three scores are below 29, and three scores are above it. Note that there are an odd number of scores in this set. It is easy to determine the median when there are an odd number of scores. Here is another example, with an even number of scores. The median will be between the two middle scores.

In this case, the median is 29.5, halfway between 29 and 30. That is, the median is a number that is not actually a score in that set. When the score set contains an odd number of scores, the median will be the middle score. When there is an even number of scores, the median will be the average of the two scores that straddle the middle of the score set.

Note one other thing about this second example. Although the one additional score (48) is quite a bit higher than the rest of the scores, the change in the median is very small. The median is not affected by outliers the way that the mean is. In fact, substitute a score of 98 instead of 48 and the median is still unchanged. The median is best used when you are concerned that outliers might be affecting the mean by making it less representative of a group of scores.

Now let's examine Ms. Dailey's classroom results. What is the median for Ms. Dailey's class on the unit test? We obtain the median by arranging the scores from lowest to highest, which we have already done in Figure 11.2. The total number of scores is 26, an even number, so we will need to identify score number 13 and score number 14, which would represent the two middle scores in this set. You can do this easily by starting at one end of the frequency table and counting the tally marks until you get to scores 13 and 14. In this case, both number 13 and number 14 are among the tally marks for score 43. The average of 43 + 43 is 43. Therefore, the median of this set of test scores is 43. Using the median as a measure of central tendency, we would say that the typical student performance is 43 correct out of 50, or 86 percent correct.

Note how the median score is higher than the mean we calculated earlier. The median is 43, while the mean is 40.7. Is this what we would 310311have predicted given the skewed data? The answer is yes. In Ms. Dailey's class, several students had unusually low scores that pulled the mean lower. The mean used all the scores, while the median did not. When the data may be skewed, consider the median as more representative of typical student performance than the mean.

Reexamine Ms. Dailey's frequency table in Figure 11.2 to see if the mean or the median appears to be more representative. The median, at 43, appears to be more typical than the mean of 40.7. A look at Figure 11.7 will confirm this. This highlights the importance of first visually examining assessment scores with a frequency table.

Mode

The third measure of central tendency is the mode, the most frequent score in a set. If you had the frequency distribution shown in Figure 11.10, the mode would be 18, the most frequent score.

In Figure 11.11, the most frequent scores are 16 and 20, with three tallies each. In this case, there are two modes. We call this a bimodal distribution (bi for “two”). We could have three, four, or more modes. We would call these multimodal distributions. This would change our picture of the typical student performance. Figure 11.12 provides histograms of a unimodal distribution and of a bimodal distribution. What do you think is the typical student performance in the bimodal distribution? Which measure of central tendency is most accurate? Or should you use several?

Figure 11.12 Histograms of Unimodal Distribution and Bimodal Distribution

311 312

Let's look at Ms. Dailey's assessment results. What was the mode of the unit test scores? Look at the frequency table again in Figure 11.2 and identify the most frequent score. In this case, the mode is 44. Like the median, the mode does not take into account all scores. It has the advantage of not being influenced by outliers to the same extent as the mean. But, be careful, the mode can occur anywhere in a set of scores and may not actually represent the most typical performance. It may also occur at several scores (such as bimodal or even multimodal). Again, a visual examination of a frequency table is in order to make sure you are not making incorrect inferences that could lead to poor decision making. Nonetheless, knowing what the most common scores are can help a teacher make an informed decision about a class as a whole and about individual students. It is possible to detect trends and patterns by looking at the mode.

Measures of Variability

So far, we have examined patterns in assessment data that can give us an idea of what typical student performances might be. These are the measures of central tendency. We might also ask, “How consistent or spread out are the student scores?” The answer could give us an idea of the variability of student learning and the overall effectiveness of our teaching. Measures of variability help inform teachers about student learning because they examine the consistency of student performances and whether scores are spread out or bunched together. We will look at two common measures: range and standard deviation.

Range

When we think about how consistent or diverse a set of assessment scores might be, we often look at the lowest score and the highest score. 312313 313314 314315These two scores indicate how wide the set of scores is. We call the difference between the highest score and lowest score the range. For example, suppose the highest student performance was 29 correct out 30 possible. The lowest score was 16 out of 30 possible. The range is simply 29 minus 16, or 13.

Resource for Your Assessment Toolkit A Quick Overview of Central Tendency Statistics

Since norm-referenced standardized testing uses central tendency statistics to make inferences and draw conclusions, it is helpful to have a quick review of some of the benefits and limitations of central tendency statistics. Sometimes it's efficient to describe a whole group of scores by giving a single number. An average summarizes a group of numbers by calculating a central tendency. There are three main types of averages: mean, median, and mode. Each of these can be used to quickly describe a set of test scores or other numbers in a group.

However, even though all three numbers provide a measure of central tendency, they must be carefully interpreted because they represent different ways of thinking about the center of a group. The following examples will help you quickly recall the issues that surround the use of means, medians, and modes.

Mean

The mean is the average you are already familiar with. It's also called the arithmetic average because it's the average you use simple arithmetic to calculate. To find the mean of a group of test scores, for example, you add up all of the scores and divide by the number of scores you have.

Let's try an example. Suppose you have the following group of ten student scores after giving a 100-point test in your class:

85 88 92 90 90 85 85 90 95 80

First, we add up the scores, giving us a total of 880. Next, we divide 880 by the number of student scores, which is 10. This gives us a mean of 88. So you can quickly describe this group of test scores by saying, “Ten students took this test and their average or mean score was 88 points.”

Advantage of the Mean: The advantage of the mean compared to the other measures of central tendency is that it takes all of the scores into account. None of the scores is left out or given any special weight in the calculation.

Disadvantage of the Mean: The mean's advantage can become a disadvantage. Because the mean takes all of the scores into account, it can be skewed or distorted by a small number of scores that are quite different from the rest, or even by just one very different score. When this happens, the number you use to describe your set of test scores may also be distorted.

Let's go back to our example and add a couple of outliers, that is, scores that are quite different from the rest. Suppose two of the scores were 0 instead of 90:

85 88 92 0 0 85 85 90 95 80

Again, we add up the scores and this time we get 700. Dividing by ten gives us a mean of 70 instead of 88. Now if we say, “Ten students took this test and their average or mean score was 70 points,” how accurately are we describing the group of test scores? Seventy points is lower than almost all of the scores and is not very accurate description of the group as a whole.

Median

The median is the number in the middle of your set of test scores after you have arranged them in order from lowest to highest. Let's go back to our first set of scores:

85 88 92 90 90 85 85 90 95 80

If we arrange them in order, this is what we have:

80 85 85 85 88 90 90 90 92 95

Then we count to the middle number—except there is no number in the middle since we have an even number of test scores. So we look at the two middle scores, which are 88 and 90, and the median will be the point halfway between them, or 89. This number is almost identical to the mean of this group of scores, which is 88.

You can use a little arithmetic to calculate the median when you have an even number of scores. Add the two middle scores (88 + 90 = 178), then divide by two to get the mean of these two numbers (178 / 2 = 89). So when you have an even number of scores, you may find yourself calculating a mean while you are on your way to finding the median!

Now let's find the median of the second example, the set of scores with two zeroes. If we put those scores in order from lowest to highest, we get:

0 0 80 85 85 85 88 90 92 95

The median again is halfway between the two center numbers, both of which are 85. But since they're the same, the median is simply 85. This median is not very close to the mean we calculated for this set (70), although it is close to the median of the previous set of scores (88). Looking at this second set of test scores, which average seems more accurate for the group as a whole, the mean of 70, which is lower than most of the scores, or the median of 85, which is close to most of the scores?

Advantage of the Median: In contrast to the mean, the median is affected very little by outliers. As a result, it tends to be a more “natural” description of the overall group performance than the mean is when outliers are present.

Disadvantage of the Median: Sometimes the median is a number that is not actually in the set of scores (like 89). Also, when there is a large number of scores, it can take a lot of time sorting the scores from smallest to largest and then counting to the middle. Of course, with the help of a computer-sort program, this problem is resolved.

Mode

The mode provides another way of looking at the central tendency of a group of scores. This type of average is the score that occurs most often in the group. To determine the mode, you simply look at all the scores and determine which score occurs most often. Returning to our example, you can see that the score of 85 and the score of 90 each occur three times. All other scores occur fewer times, so the mode is both 85 and 90. There are two modes or two central tendencies. Sometimes this is called a bi-modal distribution of scores.

80 85 85 85 88 90 90 90 92

Now let's find the mode for the set of scores with two zeroes. In this example the number 85 occurs the most often, hence the mode is simply 85. The spread or range of scores does not affect the mode. The only thing the mode represents is the score that occurs most frequently.

0 0 80 85 85 85 88 90 92 95

Advantage of the Mode: The mode is simple to determine and accurately represents the most frequent score or scores in a group.

Disadvantage of the Mode: The mode only focuses on the most frequent number or numbers in a group. The mode completely leaves out all other scores even if other scores occur quite frequently.

What is the range on Ms. Dailey's unit test? Again, look back at the frequency table in Figure 11.2, and you will see the highest score was 50, while the lowest score is 22. The range is 50 minus 22, which is 28. What does this tell us about the student performance in Ms. Dailey's class? It tells us that the scores, with a high of 50 to a low of 22, were quite spread out. Of course, this is only a limited picture of how the students performed. We only have two data points. What if the high score or the low score is an outlier—unusually high or unusually low? How might this affect our interpretation and use of the range as an indicator of how variable the scores are?

In the case of Ms. Dailey's class, we already determined that the low scores are outliers—out of the ordinary and not typical. Recall the negative skew to the data. Maybe the use of the range would lead us to conclude the scores are more spread out than they really are. A different measure of variability is in order—one that uses more of the data. In fact, what if we could use all of the assessment scores to gauge how spread out the scores are? The standard deviation is just such a measure.

Standard Deviation

The standard deviation (SD) is a measure of the average distance each individual score is from the mean. It is an indicator of how spread out the scores are around the mean. If the standard deviation is relatively small compared to the mean, then the scores are bunched together. We say that they are more homogeneous (that is, on average, the individual scores do not deviate much from the mean). On the other hand, if the standard deviation is relatively large, the scores are more heterogeneous and spread out (that is, on average, the individual scores do deviate quite a bit from the mean). It might be helpful to think of the word standard as “average” and the word deviation as “distance.” You can think of standard deviation as the average distance that individual scores are from the mean.

Figure 11.13 shows two sets of test scores. In each case the average (mean) score earned by students is 50, but the scores are not distributed in the same way. The first has a smaller standard deviation (SD = +/−3.0), the second a larger standard deviation (SD = +/−5.3). Note how the scores in the smaller standard deviation example are homogeneous—in other words, grouped together. On average, most students are clustered around the mean by plus or minus 3.0 points. Scores are more spread out and heterogeneous in the larger standard deviation example. Most students in this example are spread out from the mean by plus or minus 5.3 points. The standard deviation is one way to tell how spread out or clustered a set of scores are from the mean. This helps you as the teacher see how variable student performance is on a classroom assessment.

Figure 11.13 Two Distributions with the Same Mean but Different Standard Deviations

315 316

The formula for calculating the standard deviation appears complicated, but conceptually it is really quite simple. The formula is

where represents each individual score minus the mean, squared; and N is the number of scores that you have.

An explanation for the squaring and square root are in order. Let's assume we have the following individual scores from an assessment.

15 15 20 20 25 25

We are interested in determining the average distance of each score from the mean. Our first step is to calculate the mean. Note in Figure 11.14 that we calculate the mean in the first column by summing all the individual scores (Σ) and then dividing by the number of scores (N). In the second column, we write the mean . In the third column, we subtract each individual score from the mean.

Recall that our goal is to calculate the average distance each score is from the mean. It would seem reasonable to just calculate the average by summing all the differences (the deviations) and dividing by N. If we did that, however, we would get zero when we summed all the differences! Well, we know that the average distance is not zero, so how could that be? This happens because half of the differences between each score and the mean will be negative, and the other half will be positive. When we sum 316317them, they will add up to zero. Double-check this by adding together the differences found in the third column of Figure 11.14. You should have found that (−5) + (−5) + 0 + 0 + 5 + 5 = 0.

A partial solution to this dilemma is to square the differences. Remember, squaring a negative number gives you a positive number. In the fourth column, you will see what happens when we square the differences and add them together. We arrive at what is called the sum of the squared deviations (SS). The average sum of the squared deviations is the variance. This is simply the SS divided by N (the number of scores). We eliminated the zero, but now we have inflated deviations because of the squaring. To bring the inflated deviations back in line with the original set of scores, we take the square root of the average squared deviations.

The standard deviation (SD) for our set of scores is

317 318

You can translate this as, “The average distance of each score from the mean of 20 is 4.1.” This lets us know that, on average, the scores in this example are approximately 4 points above or below the mean. We can use the standard deviation as an indicator of how spread out the scores are around the mean of 20. Remember, the larger the SD, the more spread out the scores.

On the surface the formula seems complicated, but if you work through a few examples, it will become clear why we square the difference between each individual score and the mean and why we take the square root of the average squared deviations. Try practicing the steps outlined above on Ms. Dailey's unit test scores. You should find

Compared to the range, the standard deviation has the advantage of using all the scores in a set, so it is more likely to be representative of the spread of scores. Rather than reporting only that the scores had a range of 28, Ms. Dailey now has a reference point (the mean) and a number that tells her the average distance of the scores from that reference point. In fact, the standard deviation is often used as a unit for measuring. Ms. Dailey could use the standard deviation to answer such questions as which students scored one standard deviation (1 SD) higher than the mean or how many scores on the test were more than one standard deviation (>1 SD) below the mean.

The standard deviation does take more effort to calculate, though. The range is easier and quicker to estimate but has the disadvantage of being greatly influenced by unusually high or low scores (outliers).

So far, we have looked at several ways to visualize assessment data and at statistical concepts that can help a teacher make more informed decisions in the classroom. As a teacher, you can display your test scores in meaningful ways, using a frequency table, a histogram, or a frequency polygon. You can extract the meaning from a hodgepodge of scores by looking at measures of central tendency—the mean, the median, and the mode. And, you can gauge the variability among scores by calculating 318319the range and the standard deviation. These statistical measures are tools that help shape and guide instructional decision making in the classroom. The measures of central tendency and variability can be used to judge whether students met learning objectives and how effective instruction was.

Ask Yourself

We have demonstrated several ways in which you can interpret the performance of students individually and in groups (such as classes or grade levels). Think for a moment about the kinds of questions that parents might ask you: How is my child doing in class? How have her test scores been? Have you seen a change or growth in my child's scores? What is her average score on the assessments? Now, with an understanding of central tendency and variability of scores, what additional information might you share with parents? How could you make such information meaningful to parents who seem interested, for exam-ple, only in their child's test scores? Can you show a change in learning over time?

The Normal Curve

In our discussion of statistical concepts, we have highlighted the fact that it is good practice for teachers to graphically and pictorially represent a set of assessment scores. This is done to get an initial sense of the shape of the distribution. We noted that measures of central tendency are often different and that a set of scores may very well be skewed positively or negatively. The normal curve, or the normal distribution, is a special case. It is a theoretical, mathematically derived frequency distribution.

Look at the normal distribution in Figure 11.15. Note how the shape is symmetrical around the mean. If you draw a vertical line through the middle of the distribution, you will see that the left side of the distribution is a mirror image of the right side. Notice that the tails of each half do not actually touch the horizontal base line, that is, there never is a frequency of zero. Instead, the tails extend infinitely because the normal curve is a theoretical concept that could apply to an infinite number of observations.

In a normal distribution, all of our measures of central tendency are the same. Mean = median = mode. The average score is right in the middle. The median, by definition is also the middle score. Half of the scores are above it, and half are below it. And the mode (the most frequent score) happens to be the same score as the mean and the median. This is indeed a special situation.

The variability of scores as measured by the standard deviation also takes on special significance here. It turns out that in a normal distribution, a specific proportion of scores will be within a given distance from the mean. In this case, about two-thirds (68.26%) of the scores will be within one standard deviation below and one standard deviation above the mean (that is, +/− 1 SD). Furthermore, 95.44 percent are within plus or minus two standard deviations of the mean, and 99.72 percent are within plus or minus three standard deviations of the mean.

Because of these unique mathematical characteristics, the normal curve can be used to estimate the overall performance of a group of individuals and also to place an individual score in context with a group of scores. The normal curve is critical to the proper construction, scoring, and interpretation of norm-referenced, high-stakes assessments, which we will be discussing in Chapter 12. Measurement specialists assert that many human behaviors approximate a normal distribution. When a large enough sample of data is obtained, it is often the case that the distribution will take on characteristics of the normal curve. For example, the normal curve provides the foundation for creating, scoring, and interpreting intelligence (IQ), aptitude, and achievement assessments, such as the Stanford-Binet Intelligence Scale, the Weschler Intelligence Scale, the American College Testing Program tests (ACT), and the Scholastic Assessment Test (SAT).

320 321

Digging Deeper What Is “Normal” Behavior?

A number of social scientists and measurement specialists contend that, if our sample is large enough, natural phenomena will arrange themselves along a normal curve. Think about it for a moment: Whether we are measuring IQ, height, shoe size, or time to complete the Chicago marathon, all have extreme scores or values, but the majority of values will cluster around the middle of the curve (that is, the mean, median, and mode will all be nearly the same value). But there is an important difference between one's height and one's intelligence. Height is an absolute value that really cannot be argued. But intelligence, achievement, aptitude, and other such constructs are not absolute and are subject to error in measurement and interpretation.

So, when we are asked to describe “normal” behavior or performance on an intelligence test, for example, perhaps we should acknowledge that our measures are always imprecise and open to interpretation.

Ask Yourself

Would you expect classroom assessments to have a normal distribution? Do you think the size of most classrooms (number of students) will provide the variability needed to achieve a normal distribution? Would you be satisfied if your instructionally oriented assessments demonstrated a normal distribution? What would that say about your ability to effectively guide students to meet learning objectives or demonstrate mastery of knowledge, skills, and attitudes if only a few achieved the highest levels of performance? As you think about your classroom assessments, which measure of central tendency would be most useful?

Item Analysis

We now shift our discussion to item analysis. Item analysis is one of several ways to judge the quality of both teacher-made and published tests that use selected responses, the forms of assessment we discussed in Chapter 5. Far too often, teachers give an assessment, score it, and record the results without examining the quality of the items. Item analysis is a set of procedures designed to evaluate the quality of items that make up assessments.

321 322

Individual items on an assessment can have specific, unique characteristics. Two prominent and useful characteristics are item difficulty (how hard an item is) and item discrimination (how well an item distinguishes those who know the material well from those who do not). We will look at each from the vantage point of teacher-made classroom assessments.

Item Difficulty

An initial estimate of the quality of an item can be easily found by calculating the item difficulty. Item difficulty is simply the ratio or percentage of individuals who answered an item correctly. The item difficulty index is calculated using this formula:

item difficulty index = number of correct answers / total number of students who answered the item

The ratio of students answering an item correctly can range from 0.00 to 1.00 (or from 0 to 100% of the students answering the item correctly). The easier the item, the larger the item difficulty index. For example, if item 1 is answered correctly by 15 out of 20 students, then the item difficulty index is 15/20, which is 0.75 or 75 percent. If item 2 is answered correctly by 18 out of 20 students, then the item difficulty index is 18/20, which is 0.90 or 90 percent. Item 2 has a larger item difficulty index and is therefore an easier item.

It is fairly easy to obtain the item difficulty index for test items. After an assessment has been given, responses to each item are recorded and counted, including those that are correct and those that are incorrect. You can record the frequency of responses in the margin of the assessment next to each item and set of answers. Then calculate the difficulty index for each item based on the number of students getting that item right, divided by the number who answered that item. Remember, you may be dividing by different numbers if some students failed to answer some of the items.

Let's look at Ms. Dailey's unit test results for an example. Data from item 1 is given in Figure 11.16. Note the numbers written in the left margin. They are the number of students who selected each response.

Figure 11.16 Number of Student Responses to Item 1 on Great Depression Unit Test

It apears that for item 1, there were 22 students (out of the total of 26 who answered the item) who selected b, the correct answer. Therefore, the item difficulty index is 22 divided by 26, which is 0.85. Since 85 percent of the students answered item 1 correctly, the item difficulty index equals 0.85 or 85 percent.

Item difficulty is often used as a measure of how hard an item is. On a classroom assessment, you can gauge an item as being easy, medium, or difficult. If a number of students get an item correct, it can be inferred that the item is relatively easy. If a number of students get an item wrong, then it can be inferred that the item is difficult. A good assessment is one that 322323balances the difficulty of items to provide information about a range of student abilities and performances. All students should have an opportunity to demonstrate what they know and can do. This necessitates a range of item difficulties on an assessment. Students who are doing well overall will need the opportunity to show that with difficult items. By successfully answering difficult items, they are able to show the full extent of what they know and can do. Low-performing students will likewise need a chance to show what they know and can do. This can be achieved by including easier items on an assessment. Since the item we looked at above was the first test question, Ms. Dailey probably wanted to start off the test at a fairly easy level, to draw students in and reduce test anxiety.

When you look at item difficulty, it is also a good practice to compare the calculated item difficulty to the probability a student would guess the correct answer by chance alone. For items with only two response choices (such as true-false), the chance probability of guessing the correct answer is 1 out of 2, or 0.50 (50%). For a multiple-choice item with four options, the chance probability of guessing the correct answer is 1 out of 4, or 0.25 (25%). Clearly, the calculated item difficulty index should be greater than the probability of correctly guessing the answer.

The item difficulty index gives an estimate of how hard an item is for all students—those who performed well overall and those who performed poorly. But it would also be interesting to know how an item functions with high-performing students compared to low-performing students. If an item is working effectively, it should be answered more frequently by those students who know the content than by those who do not. This leads to the second indicator of item quality—item discrimination.

Item Discrimination

Item discrimination is defined as the degree to which an item differentiates those who have higher levels of achievement from those who have lower 323324levels of achievement. Item discrimination values range from −1.00 to +1.00. The discrimination power of an item is a measure of the ability of an item to distinguish between those students who performed well overall on a test and those who did not.

A positive discriminator is an item that is answered correctly at a higher rate by those who did well on the test compared to those who performed poorly. That is, the item is positively differentiating those who know from those who do not. Values above 0.00 indicate positive discrimination. The more positive the discriminator, the better the item is functioning in differentiating among varying levels of achievement. Such an item is thought to have more precision and thus is a more useful and effective test item.

A negative discriminator is an item that has a higher proportion of poorly performing students answering it correctly compared to those who did well overall. Such an item is, in essence, operating in the opposite direction one would expect (values below 0.00). This is undesirable. An item that is a nondiscriminator is one that does not differentiate between the high-performing and the low-performing students. This, too, is undesirable.

On published norm-referenced tests that have large numbers of students providing assessment data, the discrimination index is often estimated by calculating a correlation coefficient for the relationship between individual items and a student's overall performance on a test. But what can teachers do in the classroom? There is an alternative way to estimate the discrimination index that, while less sophisticated, is an effective approach for classroom tests. The procedure is similar to the one used earlier in calculating item difficulty. The goal is to compare the response rate of the high-performing students to the low-performing students on individual items. This is essentially the same thing as comparing the item difficulty for the high group to the item difficulty for the low group.

The first step is to identify the three groups of students in the classroom: the high-performing group, the middle-performing group, and the low-performing group. Rank order all of the tests from highest score to lowest score. If you have a typical classroom with between 20 and 30 students, you can split the class in three groups, as long as each group has roughly the same number of students.

The second step is to calculate the item difficulty index on each item for the high group and the low group. Since we are interested in seeing how well an item differentiates the high group from the low group, we do not use the middle group. (Their scores would needlessly complicate our calculations.) Now, for the high group, take the number of high-performing students who answered the item correctly and divide by the number of students in that group. The result will be the item difficulty for the high group. Do the same for the low group, and you will have the item difficulty for the low group.

324 325

The third step is simply to take the item difficulty for the high group and subtract the item difficulty for the low group. The difference is the item discrimination. Restated then, item discrimination in a classroom context is

Figure 11.17 shows an example from Ms. Dailey's classroom assessment. There are 26 students, with 9 in the high group, 8 in the middle group, and 9 in the low group. The response rates for each group to the correct answer and alternatives to item 6 are written in the left margin of the scoring key of the test.

Resource for Your Assessment Toolkit What Teachers Should Know and Be Able to Do: National Board for Professional Teaching Standards

The National Board for Professional Teaching Standards (NBPTS, 2002) has a set of five core propositions regarding what every teacher should know and be able to do. The third and fourth core propositions are directly relevant to the concepts covered in this chapter. Note in the following the role of assessing individual students as well as placing individual student performance in the context of the class as a whole. All five core propositions can be viewed online at http://www.nbpts.org/the\_standards/the\_five\_core\_propositio.

Core Proposition 3. Teachers Are Responsible for Managing and Monitoring Student Learning

Accomplished teachers create, enrich, maintain and alter instructional settings to capture and sustain the interest of their students and to make the most effective use of time. They also are adept at engaging students and adults to assist their teaching and at enlisting their colleagues’ knowledge and expertise to complement their own. Accomplished teachers command a range of generic instructional techniques, know when each is appropriate and can implement them as needed. They are as aware of ineffectual or damaging practice as they are devoted to elegant practice. They know how to engage groups of students to ensure a disciplined learning environment, and how to organize instruction to allow the schools’ goals for students to be met. They are adept at setting norms for social interaction among students and between students and teachers. They understand how to motivate students to learn and how to maintain their interest even in the face of temporary failure. Accomplished teachers can assess the progress of individual students as well as that of the class as a whole. They employ multiple methods for measuring student growth and understanding and can clearly explain student performance to parents. (pp. 3–4) [emphasis added]

Core Proposition 4. Teachers Think Systematically about Their Practice and Learn from Experience

Accomplished teachers are models of educated persons, exemplifying the virtues they seek to inspire in students—curiosity, tolerance, honesty, fairness, respect for diversity and appreciation of cultural differences—and the capacities that are prerequisites for intellectual growth: the ability to reason and take multiple perspectives, to be creative and take risks, and to adopt an experimental and problem-solving orientation. Accomplished teachers draw on their knowledge of human development, subject matter and instruction, and their understanding of their students to make principled judgments about sound practice. Their decisions are not only grounded in the literature, but also in their experience. They engage in lifelong learning which they seek to encourage in their students. Striving to strengthen their teaching, accomplished teachers critically examine their practice, seek to expand their repertoire, deepen their knowledge, sharpen their judgment and adapt their teaching to new findings, ideas and theories. (p. 4) [emphasis added]

Source: National Board for Professional Teaching Standards, 2002.

326 327

Calculating the item difficulty and item discrimination indexes requires time and effort, but it can deepen your understanding of both your students and your assessment. The results of such an analysis can provide an opportunity to increase the effectiveness of class discussion of assessment results, improve the quality of remediation (if needed), and improve the effectiveness of instruction. Knowing the quality and effectiveness of items to measure specific content and constructs will give the teacher insight into what students know and can do. During class discussion, the teacher can use the knowledge of item effectiveness to address student misunderstandings and misconceptions. There is an opportunity to seek and receive feedback on improving the quality of poorly performing items. Students will benefit from an honest attempt on the teacher's part to be fair and purposeful. The teacher has an opportunity to hone skills at test construction by receiving the feedback of item analysis and modifying items to improve their effectiveness.

Ask Yourself

Item analysis tells us how effective a particular test item is. Over time, as you refine your classroom assessments, you will find that some items are more effective and discriminating than others. How might you work with your colleagues (say, teachers of the same grade or subject) to develop a test item bank for items that have proven effective? How could you communicate the importance of such analysis to effective assessment?

Summary

• Examining sets of scores is often the first step in understanding group test scores. Frequency tables are probably the simplest arrangement and allow you to understand assessment data by identifying frequencies of scores.

• Histograms are a pictorial representation of data in the form of a bar graph in which each score is listed on the horizontal axis (the X-axis), and the tally or frequency of each score is displayed on the vertical axis (the Y-axis).

• There are three common ways to measure typical performance numerically. They are called measures of central tendency, and they summarize a set of scores.

• The mean is the arithmetic average of a set of scores and is calculated by taking all the individual scores, adding them together, and dividing by the total number of scores.

• The median is the middle score in a set of scores. To obtain the median, arrange the scores from lowest to highest and determine which score is in the middle, such that half of the scores are below it and half are above. When the score set contains an odd number of scores, the median will be the middle score. When there is an even number of scores, the median will be the average of the two scores that straddle the middle of the score set.

327 328

• The mode is the most frequent score in a set of scores.

• Measures of variability help inform teachers about the consistency of student performances and whether scores are spread out or bunched together. The two most common measures are range and standard deviation.

• Range represents the difference between the highest score and the lowest score.

• The standard deviation is a measure of the average distance each individual score is from the mean and indicates how spread out the scores are around the mean.

• Item analysis is one of several ways to judge the quality of teacher-made and published assessments. Item analysis is an empirical process and set of procedures designed to evaluate the quality of items that make up assessments.

• Two prominent and useful item analysis statistics are item difficulty (how hard or difficult an item is) and item discrimination (how well an item sorts those who know the material well overall from those who do not). Item difficulty is the ratio or percentage of individuals who answered an item correctly. Item discrimination is the degree to which an item differentiates those who did well on the test from those who did poorly on the test.

Key Terms

bimodal (311)

central tendency (306)

frequency polygon (305)

frequency table (302)

histogram (304)

item analysis (321)

item difficulty (322)

item discrimination (323)

mean (306)

median (309)

mode (311)

multimodal (311)

negatively skewed distribution (308)

normal curve (normal distribution) (319)

outlier (307)

positively skewed distribution (308)

range (315)

skewed distribution (308)

standard deviation (315)

variability (312)

For Further Discussion

1. Statistics are useful in explaining scores to others, but in what ways can they inform your teaching?

2. Students often are concerned with how well they did on a test—what grade they received. Consider the statistics that you have encountered in this chapter. In what ways do they lead you to ask how well your students performed on a test and, perhaps, how well you did in developing your test?

3. At what point would you say an individual student performed considerably below average or considerably above average compared to classmates? Would the individual score be 1 standard deviation above or below the mean? 2 standard deviations above or below mean? 3 standard deviations above or below the mean? At what point would you say an individual score is not typical (average)?

328 329

Comprehension Quiz

1. You have just scored a 50-point midterm examination in your freshman math class and calculate a mean score of 40. What does the mean score reveal about your students’ performance?

2. You look across two of your classes and find that both groups have a mean score of 40. But you notice that one group has a standard deviation of 5 and the other an SD of 10. What would you know about the performance of both groups?

3. On this same test, if you found that all of your students answered question 15 correctly, this would be an indicator of what form of item analysis? What are some things that this analysis might reveal to you?

4. Which of the three measures of central tendency is (are) influenced by extremely high or extremely low scores? Which is (are) not?

Relevant Website Resources

Practical Research, Evaluation & Assessment

http://pareonline.net/Home.htm

This online journal is supported by volunteers and presents refereed journal articles on all areas of educational assessment. There are many articles that address the use of statistical analysis in the classroom.

Testing and Evaluation Services

http://testing.wisc.edu/WhatDoThoseNumbersMean.htm

What do those numbers mean? This is a summary of test statistics and item analysis. The site offers a quick, one-page guide to interpreting the statistics you might encounter as part of the results of a classroom assessment and accompanying item analysis.

References

Collins, H. W., Johansen, J. H., & Johnson, J. A. (1969). Educational measurement and evaluation: A worktext. Glenview, IL: Scott, Foresman.

National Board for Professional Teaching Standards (NBPTS). (2002). What teachers should know and be able to do. Arlington, VA: Author.