ONE

# WHAT MAKES A DIFFERENCE?

At first sight multivariate data analysis (MDA) can appear diverse and impenetrable. However, it has two striking and reassuring features that will enable us to make sense of MDA techniques more easily than might be expected. Multivariate methods can be understood as logical extensions of simpler techniques, and further, they rely on the same small set of "moves" to achieve their ends. In the first two chapters we will explore the relatively simple statistics that later become the building blocks of multivariate techniques. As noted in the preface, the discussion in these chapters will assume no prior statistical knowledge and will attempt to provide an intuitive rather than a technical grasp. (For readers who have very little or no prior statistical knowledge, it would be helpful to review the section on key terms in the preface before proceeding.) In Chapter 3 we will examine the strategies that lie at the heart of multivariate techniques. Again, the aim is to make sense of these strategies in terms of their logic rather than the statistical technicalities that lie beneath.

In the following sections we approach the chapter title in two different ways. First we ask how differences can be quantified or, more precisely, how a single set of differences can be summarized numerically. Then we examine how the relationship between two sets of differences may be analyzed. To the extent that a systematic relationship is thereby detected we then have some evidence that one set of differences may account for the other. We will explore these two issues of quantifying and accounting for differences twice over: first in Section 1.1 for data in the form of scores and then much more briefly in

Section 1.2 for data in the form of categories. Since the main aim in the first two chapters is to expose the building blocks of MDA and to ground them firmly, we will make extensive use of small contrived examples. In Part 2, where we explore multivariate techniques as such, we will turn to examples of real-world data drawn from a variety of sources.

## 1.1 ANALYZING DATA IN THE FORM OF SCORES

### 1.1.1 Univariate Analysis: Capturing Differences in One Set of Scores

Imagine that we ask five individuals, conveniently named A, B, C, D, and E, to complete a measure of subjective well-being or happiness in which the possible score range is between 1 and 5 and a higher score indicates a higher level of well-being. Even more conveniently, imagine that each of these five individuals scores differently on the well-being measure. This situation is shown in Figure 1.1, where the individuals occupy the rungs of a score "ladder." Person A scored 5, person B scored 4, and so on. How can this set of different scores be summarized numerically to indicate just how different these individuals are from each other? In addressing this question, we are particularly interested in finding so-called summary statistics that are elegantly simple and that have the potential to be used as building blocks in more complex situations.

Even with only five cases, wondering about how to summarize all of their pairwise differences is hardly a simple beginning and in fact leads nowhere very helpful. Instead we begin by focusing on how much each person's score differs from a fixed reference point. For the purpose of developing building blocks, it turns out that the **arithmetic mean** is a particularly helpful reference point. The arithmetic mean is just the sum of all scores divided by the number of scores. Here the mean is 15/5 = 3, a number that identifies the midpoint of the scores in a way we will discuss later. So now we can re-express each person's score as the amount by which they "deviate" from the mean of 3. These **deviation scores** are shown to the right of the ladder and indicate that person A and person E are the most deviant, scoring 2 above and below the mean, respectively. Person C is totally nondeviant, but notice that often in practice there are no individual scores at the mean, which in fact may not be a possible score at all if it is not a whole number. As we make use of summary

| *Well-Being Score* | | *Deviation Score* |
|---|---|---|
| 5 | A | +2 |
| 4 | B | +1 |
| 3 | C | 0 |
| 2 | D | –1 |
| 1 | E | –2 |

**Figure 1.1** Well-Being Scores for 5 Individuals (A–E)

statistics such as the mean we leave the cases behind and shift to a different aggregate level of description. This issue of multiple levels of analysis is an important one that we will consider further at the end of Chapter 3.

So far we have simply replaced the five original scores with five deviation scores and so have yet to *summarize* differences in any way. Calculating the mean suggested that adding up scores is a useful summarizing strategy; five scores were replaced by one statistic that captured the middle of these data in one sense. Just in what sense becomes clear when we try the adding strategy with deviation scores and find that they will always sum to zero. This is because the mean has the interesting property that the deviations above and below it will always cancel out exactly. It is the fulcrum around which the scores will always balance. What to do after such a promising start? As usual we choose the option that has the most potential for acquiring useful building blocks. In the present situation the trick is to sum, not the deviation scores, but the *squared* deviation scores. Whether a number is positive or negative, multiplying it by itself results in a positive number, so the problematic canceling out of positive and negative deviations disappears. For the numbers in Figure 1.1, the sum of the squared deviation scores is $4 + 1 + 0 + 1 + 4 = 10$. This statistic is known as the **sum of squares** (shorthand for the sum of squared deviations around the mean) and is a true cornerstone of multivariate data analysis, as we will see.

The sum of squares captures the total amount of differences or variability in the data, but how might we summarize the *average* amount of difference? As with the mean, we can simply divide the total by the number of data points, that is, $10/5 = 2$. This average sum of squares is known as the **variance,** another cornerstone of statistical data analysis. Calculating it in this straightforward

way is appropriate in some circumstances, but with a slight adjustment we can arrive at a form of the variance that will serve many more purposes in more complex analyses. The adjustment involves dividing the sum of squares, not by the number of data points (5 in this case), but by the number of points that are free to take on any value. This adjusted divisor is known as the **degrees of freedom,** a notion we will encounter repeatedly but one we will not need to understand in detail. In the present case the degrees of freedom are $5 - 1 = 4$. This is because, once we have fixed the value of the mean in our calculations, the final data point is constrained to take on the value that will result in this outcome. All of the other four scores can take on any value in the range allowed by the measure, but the fifth score has no such freedom. This no doubt sounds rather mysterious, and in fact the concept of degrees of freedom really only begins to take hold when we treat cases as a sample drawn from a population—a major topic in Chapter 2. For now, the key point to note is that the variance is a fundamental measure of average difference or variability, and it always comprises a sum of squares divided by a degrees of freedom, in this case, the number of data points minus 1. The variance for the example data is therefore $10/4 = 2.5$.

In the variance we have a useful number that summarizes the average amount of differences in a set of scores. One of its limitations, though, is that by squaring scores we have moved away from the original scale. The original scores are on the 1–5 well-being scale, but the variance is in squared well-being units—a surreal and not very helpful scale. To remove this dislocation, we can simply "unsquare" or take the square root of the variance to produce a statistic called the **standard deviation.** For the present data, the square root of 2.5 is approximately 1.58. So if we want to summarize the scores in Figure 1.1 succinctly, we can simply note that there are five scores with a mean of 3 and a standard deviation of 1.58. A particular variance or standard deviation conveys useful information, but they really come into their own when used as building blocks. Before we start to build, it will be helpful to get a feeling for the features of data that influence the values of the mean, variance, and standard deviation. To do this, we turn to the three sets of data shown in Figure 1.2.

We have now extended our imaginary alphabetic sample to 10 individuals (A–J) and obtained their scores on three different measures, each of which has a possible score range of 1–5. The left-hand ladder contains their scores on a well-being measure, the middle ladder shows their scores on a measure of positive emotions (referred to as positive affect), and the right-hand ladder

| | *Well-Being* | *Positive Affect* | *Life Satisfaction* |
|---|---|---|---|
| 5 | A | A | AB |
| 4 | BC | B | CDE |
| 3 | DEFG | CDE | |
| 2 | HI | FGHIJ | FGH |
| 1 | J | | IJ |
| **Mean** | 3.00 | 2.80 | 3.00 |
| **Variance** | 1.33 | 1.07 | 2.44 |
| **SD** | 1.16 | 1.03 | 1.56 |
| **Median** | 3.00 | 2.50 | 3.00 |

**Figure 1.2** Three Sets of Scores for 10 Individuals (A–J)

contains their scores on a measure of satisfaction with life. The data are again depicted by placing individuals on the ladder rung corresponding to their score. The mean, variance, and standard deviation (SD), and another statistic called the median, appear beneath each ladder.

The well-being scores on the left-hand ladder have a mean of 3, a variance of 1.33, and a standard deviation of 1.16. Now that we have more individuals we can start to reflect on how they are distributed on the ladder, that is, the frequency distribution of their scores. The shape of the left-hand configuration is referred to as a **normal distribution,** meaning that most individuals appear in the middle of the distribution with decreasing frequencies for the higher and lower values. A further key feature of normality is that the distribution is symmetrical about its center, that is, the lower rungs are a mirror image of the upper rungs. The closer a distribution is to normal, the more helpful are the mean, variance, and standard deviation as indicators of the middle and spread of the distribution.

The distribution on the middle ladder of positive affect scores suggests that we have happened upon a group of people who are not very joyful. Their

scores have a mean of 2.8, a variance of 1.07, and a standard deviation of 1.03, all lower than those for the left-hand distribution. So, as we can see, compared with the left-hand distribution, the center has dropped down the ladder a little, and the individuals show fewer differences as they cluster more tightly on rungs 2 and 3. (Note that comparing means, variances, and standard deviations across distributions in this way can be done only if the two sets of scores are measured in the same units since all three statistics are "scale bound.") Whatever else we might say about the middle distribution, it is clearly nonnormal. Most individuals appear at the lower end of the ladder and the distribution is obviously not symmetric about its center. This asymmetry is referred to as a **skewed distribution,** more precisely a positively skewed distribution, because the scores "tail off" at the high or positive end of the scale. Were the tail pointing in the opposite direction, the distribution would be negatively skewed. Sometimes the asymmetry is due less to a continuous tail than to a few individuals whose extreme scores locate them well away from the crowd. Such scores are called **outliers,** which can influence the variance and standard deviation in a powerful way because of the magnifying effect of squaring deviation scores.

Whether the nonnormality of a distribution is due to a skewed tail or to outliers, both have a magnetic effect on the mean, dragging it away from the center of the data and rolling on to influence the values of the variance and standard deviation. The distorting effect on the mean can be seen if we compare the means for the left-hand and middle distributions with their corresponding medians. The **median** is the score that splits a sample in two, with half of the individuals above and half below. Unlike the mean, the median does not make use of the score values as such and so cannot be influenced by outliers or skewness. For the left-hand distribution, the mean and median are both 3, indicating no distortion. But in the middle distribution the mean of 2.8 has been dragged above the median of 2.5 because of the positive skew. In this constructed example the degree of skewness and its effect on the mean are not great, but in practice skewness and outliers can cause major distortions that then ripple through any analyses that have the mean at their heart.

Turning finally to the right-hand ladder of satisfaction scores, we see that the mean is 3 (as is the median), the variance is 2.44, and the standard deviation is 1.56. These figures are reassuring in that the mean clearly sits in the center of the distribution, and the variance and standard deviation show that individual differences are more pronounced than in either of the other two

distributions. However, these numbers are not to be trusted because they hide a crucial aspect of the data. The distribution is symmetrical, but it is not normal because it does not have one center but two. Put more technically, the distribution has more than one **mode** or peak; it is bimodal with peaks for the 2 and 4 score values. In practical terms this may be signaling measurement problems in the midrange or the presence of two subsamples. But for now the more important statistical point is that the mean, variance, and standard deviation are not designed to work with distributions where there are multiple modes or peaks. So, in summary, these statistics are most effective when they are applied to a normal frequency distribution, that is, one that is symmetrical about a single peak. In Figure 1.2 the normality of the left-hand distribution makes it a good candidate for these summary statistics, the skew in the middle one threatens some distortion, and the bimodality in the right-hand distribution undermines the use of the statistics completely.

Before we start to build with the blocks of the mean, sum of squares, variance, and standard deviation, two concluding comments are in order—a caveat and another way of thinking about these statistics to pave the way for things to come. It is important to appreciate that these summary statistics can be calculated and interpreted in a meaningful way only if the scores are measured on at least an **interval scale.** In other words, as we noted in the preface, the scale values must be mutually exclusive, in rank order, and equally spaced. Clearly, if the distance between, say, scores of 1 and 2 were not the same as that between scores of 4 and 5, then the arithmetic operations we have conducted would unravel. In terms of the ladder image, the rungs have not only to be fixed, but also to be fixed with equal spaces between them. We will return to this and other measurement issues in Chapter 2.

We can begin to open up another perspective on these statistics by raising the following question: If we had access to summary statistics but not the individual scores for a group of people, what would be our best guess for any given individual's score? One answer to this question is to always choose the mean score for the group. To understand why, we need to rethink the distance between any individual's score and the mean as the amount by which our mean-inspired guess has failed. Viewed in this way, the sum of squares is the total amount by which the mean "misses" individual scores. If everybody scored at the mean, there would be no misses, but obviously this is exceedingly rare. A key property of the mean is that if it is used as the reference point in calculating individual differences, it will produce a smaller sum of squares than will

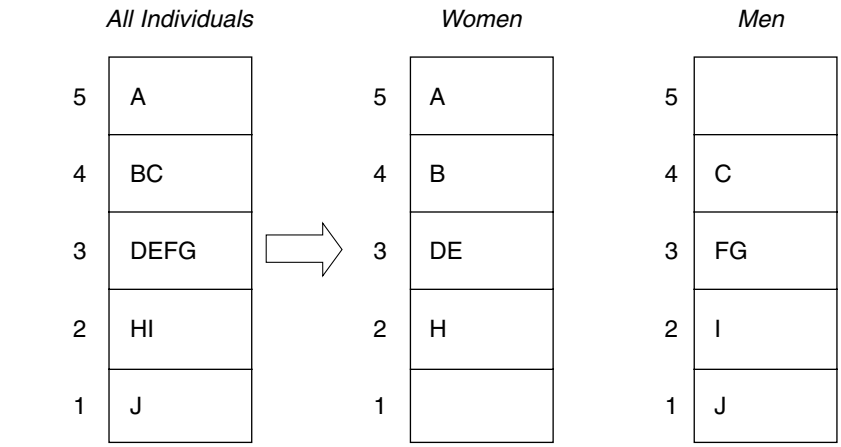
any other reference point such as the median. So the mean is a best guess of individual scores because it minimizes the miss rate, as long as we think of misses in terms of squared distances from the mean. Details aside, the key general idea to note for now is that guesses or predictions about individual scores based on averages almost always fail to some extent. The extent of this failure can be quantified using the sum of squares, variance, and standard deviation. So these statistics can be used as indicators not only of individual differences, but also of error.

### 1.1.2 Bivariate Analysis: Accounting for Score Differences With Categories

We have explored only a few building blocks so far, but already we can start to put them to work in interesting ways. Returning to the 10 individuals shown in the left-hand ladder in Figure 1.2 and reproduced on the left-hand side of Figure 1.3, we can speculate about what factors might account for differences in well-being and now actually undertake an analysis to evaluate the speculation. There is some evidence that women are likely to experience higher levels of well-being than men (Wood, Rhodes & Whelan, 1989). Is this difference evident in the data shown in Figure 1.3? Now we are treating well-being as a dependent variable and gender as an independent variable. Gender is a **categorical variable** in that it simply assigns individuals to unordered categories, two in this case. We can analyze the relationship between a dependent variable consisting of scores and an independent variable consisting of categories, using simple extensions of the statistics we now have at our disposal.

Figure 1.3 shows the well-being scores for the 10 individuals split into two groups: 5 women on the middle ladder and 5 men on the right-hand ladder. Underneath the ladders are the means, sums of squares (SS), variances (Var), and standard deviations (SD) for each grouping.

The similar variances and standard deviations indicate that there is a very similar amount of individual differences or variability within each of the three groupings. Note that the amount of variability in the women and men subgroups is identical because the two distributions happen to be mirror images of each other. But the question we really want answered is whether men and women differ in well-being as *groups,* not as individuals. The answer to this clearly lies in the *difference* in their mean scores: $3.4 - 2.6 = .8$. This difference suggests that at least in this sample of 10 people, women report higher levels




| | All Individuals | Women | Men |
|---|---|---|---|
| **Mean** | 3.00 | 3.40 | 2.60 |
| **SS** | 12.00 | 5.20 | 5.20 |
| **Var** | 1.33 | 1.30 | 1.30 |
| **SD** | 1.16 | 1.14 | 1.14 |

**Figure 1.3** Well-Being Scores for 5 Women and 5 Men

of well-being on average. This is a clear outcome, but how might we extend our simple statistics to dig deeper and open up analytic possibilities for more complex and realistic situations in the future?

The key point to note first is that we now have analyses going on at two levels: the level of individual differences and the level of group differences. Put another way, we are now quantifying **within-group** and **between-group differences.** The next step is to ask how we might combine these two sorts of information in helpful ways. It would be particularly helpful to be able to compare them and to ask whether the gender difference in well-being is greater than we would expect on the basis that individuals differ from each other anyway. In other words, is the between-group difference more than we would expect simply from within-group differences? We could explore this question at this point by considering techniques that focus only on a difference between two groups. But, as usual, we want to develop strategies that can subsequently be extended into more complex situations such as those involving more than two groups and multiple independent and dependent variables. Accordingly, we now introduce a technique called the **analysis of variance (ANOVA),** which appears in many forms throughout the realms of data analysis.

Earlier we found that a helpful way to capture individual differences was to begin with the notion of deviations from the mean and then develop this into the sum of squares and variance statistics. We can use exactly the same approach to capture differences between *groups.* Instead of focusing on mean differences as such, we calculate how far each mean deviates from the mean of the means, that is, the mean for the whole group. Looking back at the means in Figure 1.3, we see that the women's mean is .4 above the mean for the whole group and the men's mean is .4 below it. If we square each of these and sum the results, we arrive at a sum of squares that captures how different the means are. However, later we will want to compare these between-group differences with within-group differences. At the moment the two would not be comparable because one uses groups as the unit of analysis and the other uses individuals. The obvious way out of this is to convert the group calculations back to the individual level by multiplying each squared deviation score for a group by the number of cases in that group. So, the **between-groups sum of squares** would be calculated as $5(+.4^2) + 5(-.4^2) = 1.6$. This number translates the mean difference into a more flexible statistic that can be subsequently compared with individual level differences and applied to any number of means. As in the case of the sum of squares for scores, we can convert the between-groups sum of squares into a **between-groups variance.** We do this as before by dividing it by the appropriate degrees of freedom: the number of data points minus 1. Since the between-groups sum of squares is based on two means, the **between-groups degrees of freedom** is $2 - 1 = 1$. So the between-groups variance is $1.6/1 = 1.6$.

Soon we will pull all of these elements together into a coherent and hopefully satisfying pattern, but there is one more step before we do. We have discussed how to quantify individual differences for the group as a whole and the mean difference between the subgroups of women and men using appropriate sums of squares and variances. What about the differences *within* the subgroups of women and men? The subgroup sums of squares are shown under the respective ladders in Figure 1.3. These can simply be added together or pooled to produce the **within-groups sum of squares:** $5.2 + 5.2 = 10.4$. As usual, this can be converted into a **within-groups variance** by dividing by the appropriate degrees of freedom. Each group has $5 - 1$ degrees of freedom, and these are again pooled to produce the within-groups degrees of freedom of $4 + 4 = 8$. So the within-groups variance is $10.4/8 = 1.3$.

This completes all of the calculations we need to provide a comprehensive summary of the differences shown in Figure 1.3. The results of an analysis of

**Table 1.1** ANOVA Summary Table Showing the Relationship Between Gender and Well-Being

| *Source of Differences* | *Sum of Squares* | *Degrees of Freedom* | *Variance* |
|---|---|---|---|
| Between groups | 1.60 | 1 | 1.60 |
| Within groups | 10.40 | 8 | 1.30 |
| Total | 12.00 | 9 | |

variance, however simple or complex, are conventionally shown in a summary table like that in Table 1.1. This simply arranges the statistics we have calculated in a convenient pattern and so contains no new numbers.

Before we review the numbers and make use of them, a few comments on alternative terminology in this sort of summary table will be helpful. It is common for the first column to be headed "source of variability." "Source of differences" is used here instead to keep the discussion consistent, although the terms are logically equivalent. The label "within-groups" is sometimes replaced by "error," which should at least resonate with the earlier comment on variability as an indicator of failure to predict. Finally, variances in this context are usually referred to as "mean squares." This is a more informative label since it refers to the mean sum of squares, but again the "variance" label has been kept for consistency and to minimize confusion. So encountering expressions such as "error mean square" should not be a cause for confusion but an occasion for translation into the more familiar "within-groups variance."

The three rows in the summary table highlight the fact that we have analyzed the well-being differences in three different ways. In the bottom row we find the statistics that index differences in all of the cases: the total picture. This total has been split into differences that can be accounted for by being in the men's or women's groups: the between-group differences and the within-group differences, which are just individual differences. Notice that in the case of sums of squares and degrees of freedom, this split is additive, that is, the between- and within-group numbers literally add up to the total. Sometimes this is referred to as "partitioning" or dividing up the total variability into its components. This additive property will turn out to be especially valuable when we later try to make sense of complex sets of differences. Notice also that variances are not additive, which is why there is no variance entry in the "total" row. As we saw in Figure 1.3, the total variance for the whole group is

1.33, but this is not the sum of the between- and within-group variances, and so to include it in the summary table could be misleading.

Having taken the trouble to calculate these statistics, how might we use them to shed further light on the question of whether and how men and women differ in their well-being scores? Earlier it was suggested that we might compare the between- and within-group variances to see if the former was sufficient to "rise above" the latter. This can be done by dividing the between-groups variance by the within-groups variance to produce a statistic called the ***F* ratio.** In the present case this would be 1.6/1.3 = 1.23. The more this rises above 1, the more evidence we have that between-group differences are present. However, it cannot be interpreted as a direct measure of the *amount* of group difference, which is still best captured by the actual mean difference or some derivative of it. The *F* ratio is indicative of between-group differences, but we will postpone consideration of its more legitimate uses until Chapter 2. This is also why it has been omitted from the summary table, where it is usually shown in another column on the right-hand side.

We can combine elements from the summary table to produce statistics that summarize individual differences. One of these is called **eta²,** which is found by dividing the between-groups sum of squares by the total sum of squares: 1.6/12 = .133. When this is multiplied by 100, it can be interpreted as the percentage of total variability accounted for by the categorical variable. So, in the present case we conclude that gender accounts for 13.3% of the variability or individual differences in well-being scores in this sample of people. Alternatively, we could describe the situation in terms of *unexplained* variability by computing a statistic called **Wilks's lambda,** which is the ratio of the within-groups and total sums of squares. This is 10.4/12 = .867 and indicates that gender *fails* to explain 86.7% of the variability or individual differences in well-being scores. Clearly, eta² and Wilks's lambda are mirror images of each other and by definition their values always add up to 1 or 100 in percentage terms. Notice again that both of these statistics reflect individual differences. Only mean differences, or some statistic derived directly from them, convey what is happening at the group level.

This completes our introduction to some simple statistics that allow us to examine whether category differences can account for a set of score differences. By now you should already have some sense of how much analytic work can be done with a few simple building blocks. In Chapter 6 we will extend these ideas to much more complex situations where we encounter

so-called **multivariate analysis of variance.** Before that, Chapter 5 will introduce another multivariate technique called **discriminant analysis,** which is essentially another approach to multivariate analysis of variance that reverses the status of independent and dependent variables. So there we will be exploring how to account for category differences with scores and appreciating that many statistics are blind to the independent and dependent status that researchers choose for their variables. Notice, for example, how in the data we have been considering the statistics allow us to say either that gender accounts for 13.3% of the variability in well-being or that well-being accounts for 13.3% of the variability in gender. Alternatively again, we could accurately say that gender and well-being *share* 13.3% of their variability. The independent and dependent status of the variables comes from the research context, not from the statistics.

### 1.1.3 Bivariate Analysis: Accounting for Score Differences With Scores

In the previous section we conducted a simple analysis of variance to explore how far gender might account for differences in well-being. Now we return to the data on positive affect shown in Figure 1.2 and ask to what extent differences in well-being might be accounted for by differences in positive affect. We will continue to assume that positive affect was measured on an interval scale, and so now both independent and dependent variables are in the form of scores. To analyze whether score differences on a dependent variable can be accounted for by scores on an independent variable, we turn to a technique called **simple regression.** Since analysis of variance is a special case of regression, we can introduce the latter with very few new ideas or tools.

To build an initial bridge between analysis of variance and regression, it will be helpful to reconfigure and redescribe part of Figure 1.3. Figure 1.4 again shows the distribution of well-being scores for women and men separately. The distribution for the whole group has been removed, as have all of the ladder frames. The means for the women and men are shown as circles joined by a solid line. The mean for the whole group is shown as a dotted horizontal line. So nothing has changed from Figure 1.3 other than the depiction, which is now referred to as a **bivariate scatter plot** as it summarizes the relationship between two variables. As we saw earlier, at the group level the relationship can be described as the difference between the two means. These

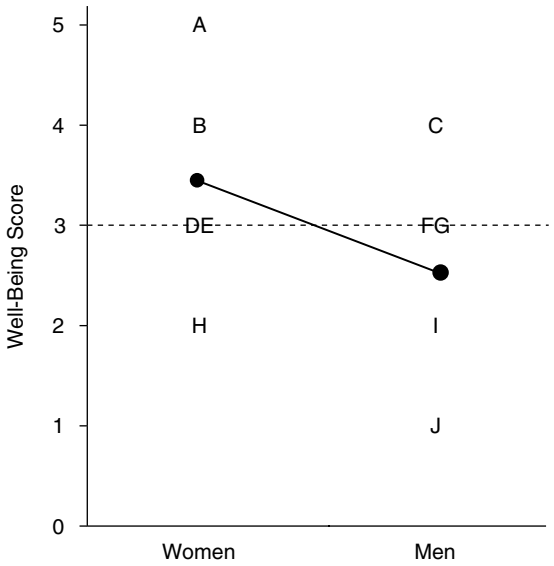



**Figure 1.4** Scatter Plot Showing the Relationship Between Gender and Well-Being

group means are also referred to as **conditional means** because their value on the dependent variable of well-being is conditional on the particular category of the independent variable. Another way to describe the group difference is in terms of the **slope** of the solid line that joins the means. This drops from 3.4 to 2.6, so the slope is –.8. The slope has a negative value because it shows a *decrease* in well-being as we travel from left to right along the horizontal axis. Finally, the group difference can also be expressed in terms of the vertical distances between the conditional means and the mean for the total group shown by the dotted line. Remember that these distances can be squared and summed to enter into the between-group sum of squares.

With these perspectives in mind, we can now focus on a regression approach to the question of whether positive affect levels might account for differences in well-being. Figure 1.5 is a scatter plot that reconfigures the well-being and positive affect data that we saw in Figure 1.2.

The dependent well-being scale appears on the vertical axis, known generally as the **Y axis,** while the independent positive affect scale appears on the horizontal, or **X, axis.** Although they are not shown as such, we now have five ladders, one for each value of the positive affect scale. In Figures 1.3 and 1.4

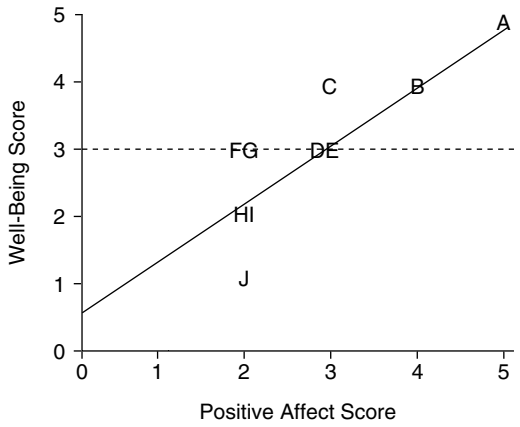



**Figure 1.5** Scatter Plot Showing the Relationship Between Positive Affect and Well-Being

we split up the total distribution of well-being scores according to the categories of the independent variable—women and men in that case. The same process has occurred in Figure 1.5, where individuals appear in vertical subgroups according to their positive affect score. This means that each letter now represents an individual's *pair* of scores: person H scored 2 on both measures, person C scored 3 on positive affect and 4 on well-being, and so on. Again the dotted horizontal line shows the mean well-being score for the whole sample. Of most importance for present purposes, the solid line again captures the relationship between the two variables at the group level and is generally known as the **regression line.** In the following we will first look more closely at this line and its interpretation as the *impact* of positive affect on well-being. Then we will bring back the analysis of variance perspective and use it to explore the relationship of these variables at the level of individual differences. Finally, we will look inside the new statistics we have encountered to see how they all make use of the same few building blocks.

### *The Regression Line*

We just noted that the solid regression line in Figure 1.5 represents the impact of positive affect on well-being in this sample of 10 individuals. To see why, it is helpful to ask first where the line comes from. A logical extension of the strategy we used in Figure 1.4 would be to draw a line that joined up the

conditional means, that is, the mean well-being scores for the subgroups of individuals at each value of the positive affect scale. However, these conditional means are not well defined. There is only one individual (A) in the 5 column, one (B) in the 4 column, and none at all in the 1 column. Moreover, in our continuing search for elegance and simplicity, it would be satisfying to summarize the relationship with a straight line—the simplifying assumption of **linearity.** So instead of joining up the conditional means, we need an alternative approach that nonetheless produces a similar outcome.

The answer is to make use of an approach based on the principle of **least squares.** This may sound daunting, but we have already encountered the principle near the beginning of this chapter. When we were summarizing one variable, we chose to use the mean as a reference point and best-guess statistic. A justification for this was that the mean resulted in a sum of squared deviations smaller than that produced by any other reference point. In other words, the mean is a desirable statistic because it obeys the principle of least squares: Its numerical value depends on minimizing the sum of squares around itself. The regression line can be thought of as a mean stretched across two dimensions. A set of data points in one dimension (one ladder) can be summarized with a point—the mean. The relationship between data points in two dimensions can be summarized with a line—the regression line. The chosen line is that which minimizes the vertical distances between itself and all of the individual data points. So the regression line cuts through the middle of the data points in the sense that it follows the straight path that produces the smallest sum of squared deviations around itself. In Figure 1.5, for example, we might be tempted to draw the regression line more steeply to get closer to individuals F, G, and C. But this would increase the distances from the other 7 individuals, especially the distant J, and result in a larger sum of squared vertical distances from the line.

Statisticians have devised computations that identify the regression line for a set of data, and we will look at some of these later. This suggests that the line can be expressed as numbers, but what are they? A straight regression line can be uniquely identified with two numbers. The first we have already introduced as the **slope** of the line. This number indicates how steeply the line goes up or down by telling us how the Y (dependent variable) value changes on its scale when the X (independent variable) value changes by 1 unit on its scale. The slope for the Figure 1.5 regression line turns out to be approximately .94. So this indicates that as the positive affect score increases by 1 unit, the happiness score increases by .94 units—almost a one-to-one relationship.

The slope identifies a family of parallel lines rather than one line in particular. To pin down the unique line for the data in Figure 1.5, we need a second number: specifically the number that tells us where the line meets the upright axis. This is known as the **Y intercept** or sometimes as the constant. Its value is approximately .37: the point on the well-being scale where the regression line meets or intercepts the Y axis in Figure 1.5. The slope and the Y intercept jointly define a unique regression line and are known generally as **regression coefficients.** They can be combined into a single **regression equation** that summarizes the relationship between any X and Y variable. Generally, this relationship is expressed as predicted Y = slope(X) + Y intercept. In words this says that the Y value for any individual can be predicted by multiplying that individual's X score by the slope and adding the product to the Y intercept. For example, the predicted well-being score for individual B who scored 4 on the positive affect measure would be (.94)(4) + .37 = 4.13.

Since the slope and Y intercept define the regression line, we can translate this idea back into graphical form. Looking at Figure 1.5, we can again make a prediction for individual B by noting what well-being score corresponds with a positive affect score of 4 according to the regression line. An imaginary line extending up from a positive affect score of 4 hits the regression line at a well-being score of just above 4, as we just calculated. This predicted score is very close to person B's actual score of 4. But notice what happens when we try the same procedure for person J with a positive affect score of 2. The predicted well-being score according to the regression line is about 2 (more precisely 2.25), but the actual score is 1. The difference between the predicted and actual scores is known as the **residual**—a statistic we will have much more to say about later. This procedure highlights the important point that we are using group summary statistics—the slope and Y intercept—to make individual predictions. Just as when we treated the mean as a basis for prediction, we are only partially successful. Some of the differences in well-being scores are explained or predicted, some are not. Again, notice how we continually move around between-group and individual levels of analysis.

The slope and Y intercept carry very different interpretative weight. The slope is the key statistic that captures the impact of an independent on a dependent variable. If someone asks us how far differences in positive affect account for differences in well-being in the present group, the slope of .94 precisely answers that question *at the group level of analysis.* As we noted, it says that a 1-unit increase in positive affect is associated with almost a 1-unit increase

in well-being on average. This slope has an implied positive sign, so it signifies that positive affect and well-being scores move up and down in concert. But a slope may have a negative value, which would mean that as the X values increased, the Y values decreased, that is, they would move in opposite directions. In general, note that a higher slope value, positive or negative, indicates a steeper regression line and greater impact. If the slope is zero, the regression line is horizontal, indicating no impact of the independent variable on the dependent variable.

The Y intercept usually has little interpretative value. To see why, note that another way to express it is as the value of Y when the X score is zero. Many measures in the behavioral and social sciences do not have a meaningful zero. Scoring a meaningful zero on a personality, attitude, or aptitude measure, for example, is rarely possible. Accordingly, the Y intercept rarely refers to an interpretable situation, though it does happen: Zero income, for example, is a perfectly meaningful, not to mention painful, situation. A final comment on the Y intercept concerns its sign. Note that a steep regression line may intercept the Y axis below the 0 point of the Y variable. In this case the Y intercept would have a negative value. This would be accurate but casts further doubt on the interpretability of the Y intercept in many if not most situations.

*ANOVA Perspective on Regression*

Earlier we noted that analysis of variance is a special case of regression. So it should come as no surprise that when we undertake a regression analysis using any respectable statistical program, the output includes an ANOVA summary table similar to the one we examined in Table 1.1. What does an ANOVA table look like when it is part of a regression analysis? Table 1.2 shows the summary table that results from the present regression analysis.

Remember that an analysis of variance splits up the total variability or differences in a dependent variable into components. Since the dependent variable of well-being has not changed, the bottom "total" rows in Tables 1.1 and 1.2 are identical. However, while the summary table in Table 1.1 showed component rows for between- and within-groups variability, the corresponding rows in Table 1.2 now refer to regression and residual sources. What exactly are these components? To answer this, we return to the "best-guess" way of thinking.

When we looked at well-being on its own, we said that the mean was our best guess for predicting individual scores, and the sum of squares (12) and

**Table 1.2** ANOVA Summary Table Showing the Relationship Between Positive Affect and Well-Being

| *Source of Differences* | *Sum of Squares* | *Degrees of Freedom* | *Variance* |
|---|---|---|---|
| Regression | 8.44 | 1 | 8.44 |
| Residual | 3.56 | 8 | 0.45 |
| Total | 12.00 | 9 | |

variance (1.33) around this mean of 3 captured the extent of our failure to predict. But now we should be able to improve our predictions if it is true that positive affect accounts for differences in well-being. To see this in action, we can refer to Figure 1.6, which is a copy of Figure 1.5 with all cases except person J removed.

If we use the mean of 3 as a basis for predicting this person's score, the error or residual would be 3 – 1 = 2: the length of the vertical line between J and the dotted horizontal line. As we noted earlier, the regression equation does a relatively poor job of predicting this individual's well-being score: a predicted score of 2.26 versus an actual score of 1 and so an error or residual of 1.26. This may be relatively poor, but the key point is that the regression information has increased our predictive power, or conversely decreased the amount of predictive error for this individual. This case-by-case analysis is interesting, but how can we combine the information into hit-and-miss indicators for the whole sample?

As we just noted, the length of the vertical line from J to the mean represents the total error we make in predicting this individual's well-being score using the mean. The segment of this imaginary line between the mean and the regression line, labeled a, represents the gain in predictive power due to our using positive affect information, that is, the gain due to regression. The remaining segment between the regression line and the individual score, labeled b, represents the amount by which we are still failing to predict, that is, the residual. The segment due to regression can be calculated for each individual. These values can then be squared and summed to arrive at the sum of squares due to regression: 8.44 in Table 1.2. This has 1 degree of freedom (defined by the number of independent variables), so the regression variance is also 8.44. Similarly, the residual segments can be squared and summed across the cases and result in a residual sum of squares of 3.56. This sum of squares has 8 degrees of freedom (defined as the number of cases minus the

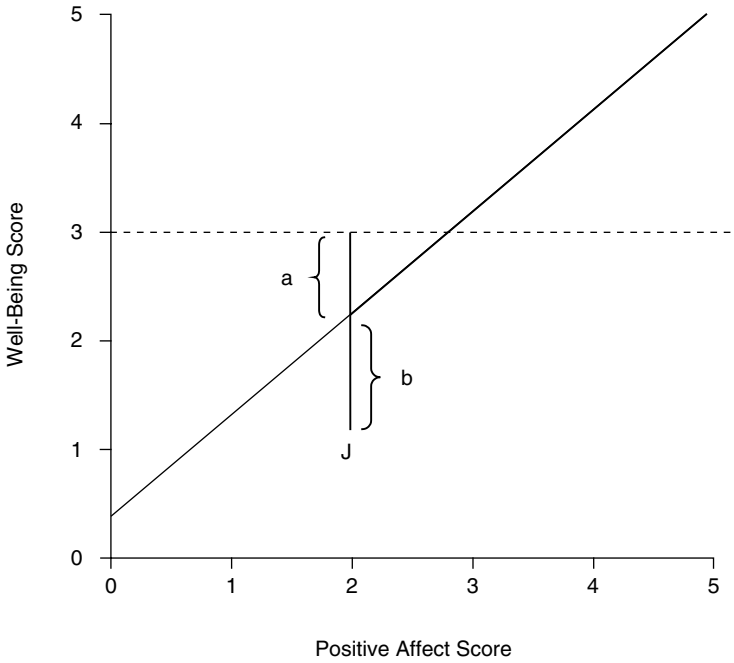



**Figure 1.6** Scatter Plot Showing the Relationship Between Positive Affect and Well-Being Showing Only Case J

number of independent variables minus 1), and therefore the residual variance is 3.56/8 = .45. It is also worth noting that in regression we often refer to the square root of the residual variance, which is known as the **standard error of estimate.** In the present data this has a value of .667 and indicates how spread out the data points are around the regression line in the form of a standard deviation (the square root of a variance).

The analysis of variance in a regression context therefore partitions the total differences or variability in the dependent variable into two components. The regression component captures differences that can be accounted for by differences in the independent variable. The residual component captures the differences that remain unaccounted for. As before, if we express these components as sums of squares, the two components literally add up to the total, as do the degrees of freedom. Also as before, we can combine elements of the table to form indices of interest. Dividing the regression variance (or mean square) by the residual variance results in the *F* ratio: 8.44/.45 = 18.8 if we

allow for rounding error. We noted earlier that the $F$ ratio is not a direct measure of the strength of a relationship and has other uses that we will explore in Chapter 2 and beyond. It is noted here mainly because it is typically included in an ANOVA summary table.

If we wish to quantify the relationship between positive affect and well-being in terms of *individual differences,* we can divide the regression sum of squares by the total sum of squares: 8.44/12 = .703. This is a direct parallel to eta$^2$, which we encountered earlier, and is known as $\boldsymbol{r^2}$ or, more formally, the **coefficient of determination.** This, too, can be multiplied by 100 and interpreted in terms of explained variability. So in these data positive affect explains 70.3% of the variability or individual differences in well-being. If we take the square root of the coefficient of determination, we produce the well-known **Pearson's correlation coefficient *r,*** which here has a value of .84. Unlike $r^2$, Pearson's $r$ can be positive or negative and so can take on values between −1 and +1. Stronger relationships move $r$ closer to +1 or −1, while a value of zero indicates the total absence of a relationship, as long as the relationship is best captured by a straight line. The value of .84 indicates a strong positive relationship such that higher scores on the positive affect measure are strongly associated with higher well-being scores.

It is important to reiterate the different types of information provided by regression and correlation coefficients, respectively. Regression coefficients, or more particularly the slope, indicate how an independent variable accounts for *group* differences in scores. The slope quantifies how group differences in the independent variable impact on group differences in the dependent variable. In contrast, correlation coefficients and their relatives show how far *individual* differences can be accounted for. Unlike regression coefficients, they express relationships in a symmetrical form. If we were to switch the independent and dependent variable status of positive affect and well-being, the regression coefficients would be different. Statistically speaking, the impact of positive affect on well-being is not the same as the impact of well-being on positive affect. However, this reversal would leave $r$ and $r^2$ unchanged. We can accurately talk about each explaining 70.3% of variability in the other or about both sharing 70.3% of their variability. As a final demonstration of the difference between regression and correlation coefficients, it is worth revisiting Figure 1.5 and noting that the regression coefficients define the regression line itself, whereas $r$ and $r^2$ convey how tightly the data points cluster around the line. A higher slope value (positive or negative) indicates a steeper line,

whereas a higher correlation value (positive or negative) indicates a tighter distribution of data points around the line.

*Inside the Coefficients*

To complete this introduction to simple regression, we will look inside the regression and correlation coefficients, to demonstrate again how much we can achieve with just a few building blocks and to show how their values are calculated. Inside we will find the mean, variance, and standard deviation. But note that none of these actually captures the bivariate *relationship* between two sets of scores; each is a so-called univariate statistic. So we need to introduce one more building block called the **covariance,** which summarizes the strength and direction of a bivariate relationship.

As the name suggests, the covariance is a close relative of the variance. To compute the variance, we calculated a deviation score for each individual, squared and summed them across individuals to produce a sum of squares, and then took the average by dividing the sum of squares by the degrees of freedom. In the bivariate situation each individual has *two* deviation scores, one for well-being and one for positive affect in the present example. The key move in developing the covariance is to *multiply* each pair of deviation scores to form so-called cross-product scores. All of these scores are then added together to produce the **sum of cross-products.** This is a direct analog of the sum of squares, but whereas the sum of squares captures all of a single variable's variability, the sum of cross-products captures all of the *covariability* between two variables. This total covariability can then be averaged by dividing through by the degrees of freedom, which again is the number of individuals minus 1. The result is the covariance, which, for the well-being and positive affect relationship, is approximately 1. The sign of the covariance, positive in this case, indicates whether the relationship is positive or negative. However, the magnitude of the covariance is not readily interpretable as it stands. Instead it becomes the key element inside regression and correlation coefficients whose magnitude can be interpreted.

The slope can simply be defined in general as the covariance of X and Y divided by the variance of X. This definition highlights how the slope indexes the strength and direction of the X-Y relationship relative to differences in X, the independent variable. So for the present data, the slope would be 1/1.07 = .94, as we noted earlier. This formula for the slope also reveals why the

magnitude of the slope would change if we reversed the status of the independent and dependent variables. Although the covariance would not change, the variance in the denominator of the slope would now be that for well-being, which is 1.33. So the slope indexing the impact of well-being on positive affect would be 1/1.33 = .75. The slopes for the impact of X on Y and Y on X are the same only when the two variables have identical variances. With the slope defined in terms of a covariance and variance, we can now use it in a simple formula for the Y intercept. The Y intercept is just the product of the X mean and the slope subtracted from the Y mean. For the present data, this would be 3 – (2.8)(.94) = .37, as we noted earlier.

Finally, what are the components of Pearson's *r* correlation coefficient? Not surprisingly, the covariance again plays the key role as it captures the strength and direction of a bivariate relationship. To convert the covariance into a correlation, we simply divide it by the product of the X and Y standard deviations. So Pearson's *r* for the well-being and positive affect relationship would be 1/(1.16)(1.03) = .84, the figure we produced earlier by another route using the sums of squares from the ANOVA table. We also noted that *r* and $r^2$ capture a bivariate relationship in a symmetric fashion: Reversing the status of X and Y has no effect. This is now further apparent in the formula using the covariance and the standard deviations. The covariance of X and Y is the same as that for Y and X, and it does not matter in which order we multiply the standard deviations: a truly symmetric statistic.

### *Overview of Section 1.1*

This completes a fairly lengthy introduction to the basic statistics used to quantify differences in one variable, and the relationship between differences in two variables, when the dependent variable is in the form of scores. Given this length, a brief overview may be helpful and serve to reiterate how these univariate and bivariate statistics are all derived from the same building blocks. In describing differences in one variable, we moved from the mean to the deviation score, to the sum of squares, to the variance via the degrees of freedom, and finally to the standard deviation. We then discussed how to discover whether category differences might account for these score differences, using an approach called the analysis of variance. This involved dividing up the total variability in the scores into between- and within-group components using sums of squares, degrees of freedom, and variances. From there we derived

three statistics: the $F$ ratio, eta$^2$, and Wilks's lambda, all using ratios of the statistics in the ANOVA summary table. However, we noted that the difference between mean scores on the dependent variable across categories remains the most fundamental way of describing the effect of the categories on the scores at the *group* level of analysis.

We then turned to simple regression, which is used to discover whether score differences in an independent variable might account for score differences in a dependent variable. The regression line, identified by its slope and Y intercept, was introduced as a way of capturing the impact of the independent variable on the dependent variable at the group level. We then saw how analysis of variance could be used to divide up the total variability in the dependent variable into predictable (regression) or unpredictable (residual) components. Again we found that elements in the ANOVA summary table could be combined to form further statistics: the $F$ ratio, the coefficient of determination ($r^2$), and Pearson's correlation coefficient ($r$). Finally, we introduced the building block of the covariance and showed how it, in conjunction with means, variances, and standard deviations, is used to build the bivariate regression and correlation coefficients.

The mean, sum of squares, degrees of freedom, variance, covariance, and standard deviation make up an immensely powerful building set. In the foregoing sections we have used them to build a first story of techniques for analyzing bivariate data, revolving around the analysis of variance and simple regression. In Part 2 of the book we will move to the higher level of multivariate analysis and see, for example, how simple regression and correlation can be extended into multiple regression (Chapter 4) and factor analysis (Chapter 7). But however complex the situation becomes, we will continue to combine these basic building blocks.

## 1.2 ANALYZING DATA IN THE FORM OF CATEGORIES

To complete this first chapter, we shift our attention to data that are all in the form of categories. Imagine that we ask a group of 10 men and a group of 10 women the simple question of whether or not they are generally happy, with the intention of finding out whether there is a gender difference. This is the same research question we asked earlier when exploring analysis of variance, but now the data have a different form. The present data come from two

**Table 1.3** Contingency Table Showing the Relationship Between Gender and Happiness

| | *Women* | *Men* | *Total* |
|---|---|---|---|
| Happy | ABCDEFGH | IJKL | |
| | 8 | 4 | 12 |
| Not Happy | MN | OPQRST | |
| | 2 | 6 | 8 |
| Total | 10 | 10 | 20 |

categorical variables each with two categories: men/women and happy/not happy. Since the measurement procedures here do not involve rank ordering or assuming equal distances between categories, all we can really do is to count the number of individuals within and across categories and develop summary statistics from there. Accordingly, this section will be considerably shorter than Section 1.1. However, when we subsequently move to multivariate analysis of categorical data in Chapters 5 and 8, it will again be apparent how powerful techniques can be developed from very simple beginnings.

Table 1.3 shows the 20 individuals, labeled A–T, distributed on a set of three ladders, each with two rungs. The left-hand ladder shows the happiness distribution for the women, and the middle ladder the distribution for the men. As noted above, all we can do with categorical data is to count frequencies, and these are shown below the letters on each rung. The right-hand ladder shows the total happiness distribution just in the form of frequencies to avoid repetition and clutter: 12 happy individuals (A–L) and 8 not happy ones (M–T). Note that, since neither variable has ordered categories, the order of rungs and of category ladders is arbitrary. This type of display is known as a **contingency table** since it shows how the distribution of a dependent variable (happiness) is contingent on the distribution of an independent variable (gender). What have been referred to as "rungs" are more conventionally known as cells. The frequencies in the four cells within the table are called the **conditional frequencies,** and those around the edges that show the total distributions for the two variables are called the **marginal frequencies.** The bottom right-hand cell contains the total frequency, which of course is equal to the number of individuals.

Given the limited form of the data, how might we quantify the distribution of a single variable such as the dependent happiness variable, shown in the right-hand column? Neither the mean nor the median would be a meaningful

summary of the middle of the distribution. Instead we use the **mode,** which is simply the highest frequency, that is, 12. So the modal category is the happy one; the most common status for this group of people is to be happy. To capture differences or variability, we can make use of ratios of frequencies in two ways. Dividing a marginal cell frequency by the total and multiplying it by 100 provides a difference statistic in the form of a percentage. So 100(12/20) = 60% of the group are happy, while 100(8/20) = 40% are not. These relative frequencies are often treated as *probabilities,* for example, a case has a 60% chance of being happy. Alternatively, we can form a ratio of marginal cell frequencies to produce an **odds** statistic. The odds of being happy in this group are 12/8 = 1.5; conversely, the odds of not being happy are 8/12 = .67. Notice that an odds of 1 indicates an equal split and therefore maximum variability, as in the case of the gender variable. The further the odds deviate above or below 1, the less the variability or difference on that variable.

One reason for introducing the odds statistic is that it can easily be developed into a useful bivariate statistic that captures the relationship between two categorical variables. If we look inside the table, we can see that the so-called **conditional odds** of being happy for women are 8/2 = 4, and the conditional odds for men are 4/6 = .67. If we divide one of these odds by the other, we arrive at the **odds ratio.** This will be 4/.67 = 5.97, which says that women are nearly 6 times as likely to be happy than not, compared with men in this group. This method of capturing the relationship between categorical differences has become a commonplace of public health messages. Statements such as "smokers are twice as likely to suffer a heart attack as nonsmokers" are based on the calculation of odds ratios. Unlike regression and correlation coefficients, where the absence of a relationship is indicated by a value of zero, an absent relationship between two categorical variables results in an odds ratio of 1. So the further the odds ratio deviates above or below 1, the stronger the relationship.

In Subsection 1.1.2, on analysis of variance, we briefly encountered a statistic called the *F* ratio. This was crudely described as a way of indexing the extent to which group differences on a dependent variable are greater than what might be expected on the basis of individual differences. We can take a similar approach with wholly categorical data, using a statistic called **chi².** This statistic can be used to quantify how much greater the group differences we see in the conditional frequencies in Table 1.3 are than we would expect given the individual differences evident in the marginal frequencies. Notice that this is equivalent to asking how far the odds ratio differs from 1, or to asking whether

there is a relationship between gender and happiness in these data. All other things being equal, the value of $chi^2$ increases with the strength of the relationship, and in the present case it has an approximate value of 3.33. However, as with the *F* ratio, $chi^2$ is not a direct measure of the strength of a relationship. Its true function will become clearer in Chapter 2 as a so-called test statistic. For now, it is sufficient to note that it can be transformed to arrive at a correlation coefficient called the **phi coefficient.** The transformation involves dividing $chi^2$ by the total frequency and then taking the square root: The square root of 3.33/20 = .41. This turns out to be a special version of Pearson's *r,* which can be used to quantify the relationship between two dichotomous variables, that is, each having two categories. Accordingly, we can report that gender and happiness are correlated .41 in the present group, or we can square phi and multiply it by 100, as we did for $r^2$, and say that gender explains about 16.8% of the variability in happiness.

Much more could be said about the nature of $chi^2$, the analysis of contingency tables with more than 4 cells, and other types of correlation coefficients for categorical data. However, the ideas and statistics we have briefly discussed are sufficient as a foundation for later chapters. As noted earlier, we will build directly on this foundation in Chapter 5 when we encounter logistic regression, and especially in Chapter 8 when we explore log-linear analysis in general to see how it can be used to analyze contingency tables with more than two categorical variables.

## 1.3 FURTHER READING

Good introductions to basic data analysis can be found in Rowntree (2003) and Rosnow and Rosenthal (2001). More extensive treatments of univariate and bivariate analysis techniques can be found in Hays (1994) and Rosenthal and Rosnow (1991). Good, focused introductions to ANOVA and simple regression are available in Keppel, Saufley, and Tokunaga (1993) and Darlington (1990), respectively.