INVESTIGATION TOOLS

DETECTING SOCIAL FACTS

While the individual man is an insoluble puzzle, in the aggregate he becomes a mathematical certainty. You can, for example, never foretell what any one person will do, but you can say with precision what an average member will be up to. Individuals vary, but percentages remain constant.

Sherlock Holmes, in
Sir Arthur Conan Doyle’s
The Sign of Four

Emile Durkheim (1858-1917) is regarded as a founder of modern sociology. To him, two issues were central to sociology: establishing a scientific basis to study society and diagnosing the breakdown of modern society. He traced this breakdown to the declining capacity of groups such as the family, church, and community to bond individuals to the social order. As a rule, the more the individual is attached to groups, the more the individual can draw upon the groups for support and strength. Detachment for the individual means personal isolation and anomie. Isolated and lonely individuals find adversity increasingly frustrating and life emptied of purpose. As their despair deepens, suicide correspondingly begins to seem an increasingly reasonable alternative. Durkheim predicted, for example, that those who are married (more attached) will have lower suicide rates than those who are divorced (less attached). Durkheim investigated these and other ideas with percentage and rate tables. In the following section, we will see how Durkheim used tables for his research, how tables are organized, how to read one-and two-variable tables, and, lastly, how to infer causal connections.

The Organization of Tables

A good table quickly tells you what it is trying to communicate. It does this by including the following elements: title, notes, headings, stub, and cells. Table 1 on the next page illustrates how these elements appear in a well-organized table.

1. Title The title introduces the reader to the subject the table is presenting. In Table 1, the title states that the table is explaining suicide rates for selected countries (the assumed effect) by age and sex (assumed causes).

2. Headnote This information immediately follows the title and is essential to a correct interpretation of the statistics that are presented in the table. Our headnote tells us that the definition of suicide that the table uses comes from the International Classification of Diseases (ICD). The LCD definition of suicide includes lethal injuries that are either directly or indirectly self-inflicted.

3. Headings and Stub The table is divided with lines, called rules, into blocks of columns and rows. The column headings in Table I identify the seven countries and years being described. The stub identifies the information in the rows. In this table, it consists of the suicide rate by age group for males and females.

4. Cells The cells provide more specific information. These are found in the stub where the rows and columns intersect. To interpret cells, we need to know whether they consist of raw or transformed numbers. Raw number tables may or may not specify whether the cells contain simple frequency counts. Tables that have transformed numbers into averages, indices, percents, rates, or ratios must identify the values in the cells. The unit indicator (just below the table title) provides this information. In Table 1, the unit indicator tells us that the numbers given are numbers of suicides per 100,000 population.

5. Footnote This information appears below the bottom table rule. In Table 1 the asterisk that follows “United States 1989” refers us to the foot of the table. This note indicates that the original table can be found in The Statistical Abstract of the United States, 1994, on page 859.
### Table 1: Suicide Rates for Selected Countries by Sex and Age Group

(Rates per 100,000 population)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MALE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>21.0</td>
<td>34.6</td>
<td>30.0</td>
<td>11.1</td>
<td>12.3</td>
<td>12.1</td>
<td>19.9</td>
</tr>
<tr>
<td>15—24 yrs. old</td>
<td>27.8</td>
<td>29.7</td>
<td>28.0</td>
<td>9.9</td>
<td>15.8</td>
<td>17.2</td>
<td>24.3</td>
</tr>
<tr>
<td>25—34 yrs. old</td>
<td>26.0</td>
<td>42.8</td>
<td>41.5</td>
<td>9.2</td>
<td>16.2</td>
<td>18.7</td>
<td>22.8</td>
</tr>
<tr>
<td>35—44 yrs. old</td>
<td>24.4</td>
<td>36.6</td>
<td>44.4</td>
<td>13.5</td>
<td>14.8</td>
<td>17.3</td>
<td>22.4</td>
</tr>
<tr>
<td>45—64 yrs. old</td>
<td>23.6</td>
<td>48.2</td>
<td>46.2</td>
<td>17.1</td>
<td>16.1</td>
<td>12.8</td>
<td>24.6</td>
</tr>
<tr>
<td>65—74 yrs. old</td>
<td>27.7</td>
<td>64.0</td>
<td>41.5</td>
<td>25.0</td>
<td>15.4</td>
<td>10.8</td>
<td>33.0</td>
</tr>
<tr>
<td>75 and older</td>
<td>39.8</td>
<td>123.6</td>
<td>69.4</td>
<td>43.6</td>
<td>34.0</td>
<td>17.4</td>
<td>54.2</td>
</tr>
<tr>
<td><strong>FEMALE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.6</td>
<td>11.6</td>
<td>15.1</td>
<td>4.1</td>
<td>7.2</td>
<td>3.4</td>
<td>4.8</td>
</tr>
<tr>
<td>15—24 yrs. old</td>
<td>4.5</td>
<td>6.1</td>
<td>3.6</td>
<td>1.6</td>
<td>3.6</td>
<td>2.2</td>
<td>4.2</td>
</tr>
<tr>
<td>25—34 yrs. old</td>
<td>7.2</td>
<td>7.6</td>
<td>8.1</td>
<td>3.2</td>
<td>7.2</td>
<td>3.2</td>
<td>5.6</td>
</tr>
<tr>
<td>35—44 yrs. old</td>
<td>7.5</td>
<td>12.4</td>
<td>14.4</td>
<td>3.5</td>
<td>8.7</td>
<td>4.2</td>
<td>6.6</td>
</tr>
<tr>
<td>45—54 yrs. old</td>
<td>8.2</td>
<td>16.3</td>
<td>21.7</td>
<td>5.1</td>
<td>10.1</td>
<td>4.6</td>
<td>7.3</td>
</tr>
<tr>
<td>55—64 yrs. old</td>
<td>8.7</td>
<td>16.1</td>
<td>30.4</td>
<td>6.7</td>
<td>12.1</td>
<td>5.3</td>
<td>7.3</td>
</tr>
<tr>
<td>65—74 yrs. old</td>
<td>7.4</td>
<td>19.2</td>
<td>32.5</td>
<td>8.5</td>
<td>9.4</td>
<td>5.2</td>
<td>5.9</td>
</tr>
<tr>
<td>75 and older</td>
<td>10.0</td>
<td>26.7</td>
<td>28.4</td>
<td>9.5</td>
<td>14.9</td>
<td>5.6</td>
<td>5.9</td>
</tr>
</tbody>
</table>


### Table Reading

After discovering the table’s organization, the interpretation of its information can begin. With Table I we can consider the question of what, if any, relationship exists between suicide and sex and age. Our analysis of the table begins by noting that overall suicide rates have a range of 122.0 (from 1.6 for Italian females aged 15—24 to 123.6 for Austrian men 75 years of age and older). Both variables—sex and age—play central roles in understanding the variation in suicide rates. The ranges for men (22.8, or 34.6 minus 11.2) and women (11.0, or 15.1 minus 4.1) reveal a relationship between sex and suicide. Age, however, has an even larger range of 97.9 (25.7 for Austrian men aged 15—24 versus 123.6 for Austrian men aged 75 and older), indicating that age plays an important part in explaining suicide. In addition to the connections between age and sex, it is equally clear that great differences in suicide rates exist in the seven countries. The suicide rates of Austria, for example, are three to four times as large as those of the United Kingdom. This difference tends to hold true across both age and sex boundaries.

### Conclusions

Our study of Table I allows several conclusions: men have much higher suicide rates than women; suicide rates generally increase with age for men but not necessarily for women; and suicide rates vary greatly from one country to another. We may conclude that a very important connection exists between suicide and age, sex, and country.
One- and Two-Variable Tables

All tables try to describe or explain variables. A variable is simply a factor or phenomenon (for example, religion) that has more than one value. Variables must be divided into values that may be qualities (such as Roman Catholic or Protestant) or quantities (such as ranks or numbers). A one-variable table or array merely presents those values and the frequency of each. Table 2 is a one-variable table.

Table 2: SELECTED CAUSES OF DEATH PER 100,000 IN THE UNITED STATES, 1991

<table>
<thead>
<tr>
<th>Cause of Death</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homicide</td>
<td>10.5</td>
</tr>
<tr>
<td>Suicide</td>
<td>12.2</td>
</tr>
<tr>
<td>Accidents</td>
<td>35.4</td>
</tr>
</tbody>
</table>


To read Table 2 most effectively, identify the variable presented. In this case, the variable is Cause of Death or Selected Cause of Death. The same variable may be given different names. Do not let this mislead you. The variable is measured by an indicator. An indicator indicates the presence (or, better, the degree of presence) of a given variable. Cause of Death might be indicated by the entry on death certificates. While such entries may not always be perfectly accurate, they are probably valid (that is, actually indicating what we think they indicate) and reliable (giving the same value in instances that are the same and different values in instances that are different) enough for our purposes.

Be careful not to confuse values with variables. Remember that all variables must vary; they must have at least two values (otherwise they would not be variables, but constants). In Table 2, the variable Cause of Death has three values—Homicide, Suicide, and Accidents. These values are nominal; that is, they have names instead of quantities. Values often are quantities like 1, 2, 3, or 4. A sequence of values, such as 1st, 2nd, 3rd, or 4th, is called a rank order.

Having identified the variable and the values, we then want to know what the table tells us about the variable. To learn that, we look at the figures listed next to the values. We see that next to the value Homicide, the rate is 10.5. That means that there were 10.5 murders in the United States in a single year for every 100,000 persons. Why is this figure used? Why not simply list the total number of murders in the United States? The reason is that using a rate (which this is) or a percentage—a kind of rate that always is given per 100—the number out of 100) allows us to compare frequencies in different-sized groups, samples, or nations without being misled. For example, there are hundreds of murders (and suicides) each year in a large city like New York City, while there are only a few in most small towns. Yet this does not necessarily mean that the small town is safer than the big city. The rates (or percentages) might be the same in both places. Therefore, we use rates or percentages so that we can know whether the frequencies of phenomena are really similar or different without being misled by the difference in the size of the population or sample from which our data are drawn. Of course, more people die from all causes in large cities than in small towns. But that does not mean that the death rate is necessarily greater.

In sociology, percentages are the most commonly used form of rate. If a percentage had been used in Table 2, the right column heading would have read Percentage rather than Frequency. Percentages could have been used in this table, but because percentage means number per 100, the resulting figures would have been inconvenient to use because they would involve so many decimal places (for example, expressed as a percentage, the homicide rate of 10.5 per 100,000 would be written 0.0105%).

If we compare the frequencies shown in the table, we learn that in the United States a person is about twice as likely to die in an accident as to be killed intentionally through homicide or suicide (35.4 is about twice the total of 10.5 + 12.2). Furthermore, if a person is killed intentionally, it is more likely to be a case of suicide than a homicide.
(the homicide rate of 10.5 is less than the suicide rate of 12.2). This fact may surprise some people, since we do not often realize that we are in greater danger from ourselves than from others. Actually, if we put this information together with other data (not shown in this table), we see that most murder victims knew their murderer as either a friend, an acquaintance, or a relative before the crime. We must conclude that we are fairly unlikely to be killed by a stranger.

We can learn, as you can see, quite a bit from a one-variable table. If a detective who was familiar with these data discovered a person who had been shot to death, he would first ask whether the shooting could have been an accident (most likely) or a suicide (next most likely) before suspecting murder. And with the additional information mentioned above, even if it were it a murder, the detective would be well advised to check the victim’s friends and relatives before looking for a homicidal maniac who was a stranger to the victim. This method of investigation would be the one most likely to solve the case quickly and efficiently, despite what some popular murder stories may have led us to believe.

Two-Variable Tables
As useful as one-variable tables are, sociologists use them only marginally in their work. We are not interested in frequencies alone. We want to know how variables relate to one another. We seek associations and, ultimately, causes of social behavior. But just studying a single variable will not allow us to find associations or causes: it will show us only results or effects). A univariate (one-variable) table such as Table 2 tells us only that people do kill themselves and others, and how frequently. Sociologists, however, usually want to know why—not only in this area of behavior but in all human action.

To learn why, to seek associations and causes, we need a table showing at least two variables. This kind of table (a bivariate table) in Table 3.

<table>
<thead>
<tr>
<th>Religion</th>
<th>Suicide</th>
<th>Homicide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protestant</td>
<td>326.3</td>
<td>3.8</td>
</tr>
<tr>
<td>Catholic</td>
<td>86.7</td>
<td>32.1</td>
</tr>
</tbody>
</table>

Note: Reprinted with permission of The Free Press, a division of Simon & Schuster, from Emile Durkheim, Suicide (pp. 154, 353), translated by John A. Spaulding and George Simpson. Copyright © 1951, copyright renewed 1979 by The Free Press.

As before, the first questions to ask about this table are, how many variables are presented and what are they? Then one should ask how many values each variable has and what the values are. In Table 3, one variable is Cause of Death, as in Table 2. How many values does it have in Table 3? Not three as in Table 2, but only two. Accidental deaths have been omitted. Investigators often omit possible values to focus on those that are more important to the questions they are asking. Here we are focusing on intentional killing—suicide and homicide. The title of this table is ‘Causes of Death per million by Religion.’ The word by often links variables in a title. The second variable is Religion. How many values does it have? Again, the answer is two: Protestant and Catholic.

Table 3 is the simplest kind of two-variable table—a “two-by-two” table. It is called two-by-two because each of the two variables has two values. Depending on the number of variables and the number of values each variable has, tables can become very complex. A two-by-two table has four cells. The intersection of Suicide and Protestant gives us the rate of suicide among Protestant Europeans. In other words, in the upper left cell of this table one finds the rate 326.3. Be sure you can find all four cells and understand the figures in them before reading further. (Note that now we are using rates per 1,000,000 population)

Association

Discovering Relationships with the Diagonal Rule
Our main interest in looking at Table 3 is to discover whether the two variables are related or associated. What tends to go with what? If we knew the value of one variable, could we guess with a better-than-even chance of being
right the value of the other variable? If so, we would be well on our way to understanding (and predicting) this sort of human behavior.

To see whether a table shows an association between variables, we compare the figures (usually rates or percentages) in the cells of the table. Table 3 organizes information so we can test Durkheim’s belief that there is an association between religion and suicide. At the intersection of Protestant and Suicide, we note the rate of 326.3 per million. At the intersection of Catholic and Homicide, we find 32.1 (the lower right cell). In the upper right cell (Homicide and Protestant) and in the lower left cell (Catholic and Suicide) we find 3.8 and 86.7 respectively. How are we to interpret these figures? It seems that Protestants are more likely to commit suicide than are Catholics, whereas Catholics are more likely to be murdered than are Protestants. We would say that the variables, Cause of Death and Religion, are associated. We know this because the numbers that make up the greatest proportion of each column form a diagonal set of cells. In this table, the numbers in the upper left and lower right cells are each the largest in their columns, and they can be linked by a diagonal line. This rule of thumb, called the “diagonal rule,” is one handy way of seeing whether variables in a table are associated—that is, tend to “go together.” Put into other words, the diagonal rule says that in a two-variable table, if the proportions (percentages, rates, or absolute numbers) on one diagonal are much greater than the proportions on the other diagonal, then the variables are probably associated with each other.

In Table 3, we have found an association between religion and cause of death. We know this by following the diagonal rule. This rule is not a precise measure of association, but it does give us a first approximation—an idea of what goes with what. (More exact measures of association, discussed later, can also be used.) In this case, we would say that religion seems to be somehow related to certain kinds of behavior—murder and suicide. Protestant Europeans were found to have higher suicide rates than Catholic Europeans. Religion and suicide seem to be associated in some way. (Note that having shown only an association, we cannot speak accurately of cause.) Now, you might ask whether tables always show associations between variables. No, they do not. But what would a table showing no association be like? The answer is that it would usually be the opposite of Table 3. Instead of numbers piling up on the diagonal, the cells would have very similar numbers in the rows and/or columns. Table 4 is an example of a “no association” table.

### Table 4: Variations Over Time of the Rate of Mortality by Suicide and the Rate of General Mortality

<table>
<thead>
<tr>
<th>Year</th>
<th>Suicides (Per 100,000)</th>
<th>General Mortality (Per 1,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1849—1855</td>
<td>10.1</td>
<td>24.1</td>
</tr>
<tr>
<td>1856—1860</td>
<td>11.2</td>
<td>23.2</td>
</tr>
</tbody>
</table>

Note: Reprinted with permission of The Free Press, a division of Simon & Schuster, from Emile Durkheim, *Suicide* (p. 49, Table 1b translated by John A. Spaulding and George Simpson. Copyright © 1951, copyright renewed 1979 by The Free Press.

There is no clustering on the diagonal in Table 4. Instead, the cells show rates that are much the same. There were 10—11 suicides (per 100,000 people) and 23—24 deaths (per 1,000 people) during the time from 1849 to 1860. Here we would say that there is no association between time period and cause of death. By knowing the time period, you gain little ability to predict either the suicide or death rate. One could do about as well by flipping a coin—pure guess. Although there are other table distributions that also show no association, this kind is probably the most common. (If one does much social research, one realizes just how common findings of no association really are!)

Durkheim turned his “no finding” into an important discovery. In a more detailed study of these rates over time, he found mortality, as stable as it was, fluctuated more than suicide did. Durkheim’s discovery suggested that some feature of society was responsible for the remarkable stability of suicide rates. He understood that the individual plays a part in the decision to commit suicide, but individuals cannot determine the suicide rates across generations. He concluded that suicide rates could not be understood as a psychological fact. They must be understood as a new order of fact—as social facts. Social facts (such as the rates of crime, birth, literacy, and poverty) cannot be explained by psychological facts but only by other social facts (such as religion, family, industry, or urban life). Thus, as Durkheim demonstrated, scientific results that report no findings can be just as important as those that report significant findings.

With real data, one is unlikely to find such obvious cases as shown in Tables 3 and 4. But the general
principles of table reading apply: the more the numbers tend to pile up on the diagonal, the greater the level of association. The more evenly the numbers are (distributed among the cells, the less the association.

The same principles of table reading apply to tables that are larger than two-by-two. If there are more cells, but still only two variables, one can apply the diagonal rule. For example, if in Table 4 we had included the value Accidental Deaths under Mortality Rate and added 1870—1879 as another time period under investigation, we would have produced a table with nine cells (a three-by-three table). The two variables would then have had three values each. If we had not included Accidental Deaths but had included 1870—1879, we would produce a table with six cells. The structure of tables with two variables can grow rather complex when the number of values shown is increased. Nevertheless, as long as there are only two variables, the principles of reading are much the same.

Three-variable tables are more complex still, but most of them are really a set of two-variable tables. For example, we might show cause of death by time period and religion. Then we might show tables like those in Table 3 and 4, but one table would be only Protestants, while the other would be only Catholics. Such three-variable tables are common when one wishes to control for other variables. But remember that each component two-variable table may be read as we have described.

The principles of reading two-variable tables may be presented as follows:

<table>
<thead>
<tr>
<th>LARGE</th>
<th>smaller</th>
</tr>
</thead>
<tbody>
<tr>
<td>smaller</td>
<td>LARGER</td>
</tr>
</tbody>
</table>

This would show an association between the variables.

<table>
<thead>
<tr>
<th>smaller</th>
<th>LARGER</th>
</tr>
</thead>
<tbody>
<tr>
<td>LARGER</td>
<td>smaller</td>
</tr>
</tbody>
</table>

This would also show an association between the variables.

<table>
<thead>
<tr>
<th>about the same</th>
<th>about the same</th>
</tr>
</thead>
<tbody>
<tr>
<td>about the same</td>
<td>about the same</td>
</tr>
</tbody>
</table>

This would show no association between the variables.

Table 5 introduces a new variable, Anxiety Level. This variable could be indicated by a series of questions (indicators) forming an anxiety index. (A typical question might be: “How worried are you about the future?”) Such an index could have many values from lowest to highest. But for the sake of simplicity, we will present this variable “dichotomously”-having only two values (High and Low). If one reduces the number of values in a variable by combining values together, This is called **collapsing**

Collapsing causes the loss of some information but may make the table easier to understand.

Note that Table 5 uses percentages that are based on row totals. The actual (“absolute”) numbers for each cell are given in parentheses under the percentages. This is a very common practice. The total number studied or sampled is called the “N” of the study or table. Here, it is 1,500. (Remember that these are fictitious data.)

---

2 **MY NOTE: WE WILL BE REFERING TO THE COLLAPSING PROCESS AS ‘RECODING’.**
Table 5: LEVEL OF ANXIETY BY RELIGION*3

<table>
<thead>
<tr>
<th>Religion</th>
<th>Anxiety Level</th>
<th>High</th>
<th>Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catholic</td>
<td></td>
<td>25%</td>
<td>75%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(175)</td>
<td>(525)</td>
<td>(700)</td>
</tr>
<tr>
<td>Protestant</td>
<td></td>
<td>65%</td>
<td>35%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(520)</td>
<td>(280)</td>
<td>(800)</td>
</tr>
</tbody>
</table>

N = 1,500

*Fictitious data

Once we have located the variables and the variables’ values as we did before, we can ask if this two-variable table shows an association between the variables. Apply the diagonal rule. Are the numbers (percentages) greater on one diagonal (not row or column) than on the other diagonal? In the lower left we find the figure 65%. In the upper right we find 75%. Note that it does not matter whether the percentages are based on the row totals, column totals, or total N: the diagonal rule can still be used as a guide. In the other diagonal set of cells, we read 25% and 35%. There is indeed a piling up of these cases on the first diagonal. The variables are associated. That means, in terms of substance, that Catholics tend to be less anxious than Protestants. Although we note that 25% of the Catholics and 35% of the Protestants do not follow this tendency, that does not mean that the tendency or association does not exist. There are few perfect associations to be found in science.

We could interpret our finding to mean either that anxious people tend to become Protestants or that Protestants have a greater likelihood of becoming anxious. Or there could be some combination of these interpretations. How can we know which interpretation is the correct one? In addition to association, we must also know the time order of the variables. A cause4 (what sociologists call an independent variable) must always come earlier in time than an effect (called a dependent variable). Time order is usually easy to establish when you compare ascribed or inherited characteristics (such as race, sex, or age) with achieved or earned characteristics (for example, social class, education, occupation, or crime). It would not make sense to say that education causes people to be male or female, since one is male or female from birth; sex is determined prior to education. Time order becomes more difficult to determine in situations like that which exists between religion and anxiety. In such cases we must often, like Durkheim, rely upon a theory to establish causal order.

If we could show that religion does come first, we could propose that the religion is the cause of anxiety. We might establish prior occurrence by asking in what religion subjects were raised. Then we might assume that a person’s current anxiety level comes after the religion acquired in childhood. If this were our interpretation (and we would still need to control for other factors or variables before we could accurately speak of causes), how might it relate to the suicide rates shown earlier?

In his study entitled *Suicide* (1897/1951) Durkheim reasoned as follows: Suicide rates are functions of unrelieved anxieties and stresses to which persons are subjected. Catholics have lower levels of anxiety than do Protestants; therefore, there will be lower suicide rates among Catholics than among Protestants. This is exactly what he found to be true. Protestants do have higher suicide rates. If Table 5 were not fictitious, it would lend further support to Durkheim’s hypothesis relating religion, anxiety, and suicide. But why should Protestants experience greater anxiety? To understand that, Durkheim introduced another variable, social cohesion. The anxiety may come from lower social cohesion (social support) among Protestants. Protestantism puts more emphasis on the individual’s responsibility before God for his or her sins. On the other hand, the Catholic Church mediates to a greater extent between God and man. It could be, then, that the Catholic facing stress does not feel quite so alone or individually culpable as does the Protestant. The Catholic has greater social support in dealing with stress than does the Protestant, who is more likely to put the blame for difficulty directly on him- or herself. If this line of reasoning is correct, one could understand higher anxiety among Protestants’ leading to a higher suicide rate. The causal chain

---

3 MY NOTE: PLEASE NOTE THAT IN THIS CLASS WE WILL BE USING COLUMN PERCENTAGES ONLY

4 MY NOTE: IN THIS COURSE WE WILL NOT USE THE TERM ‘CAUSE’.
would be as follows: religion leads to greater or less social cohesion, which leads to greater or less unrelieved anxiety, resulting in different suicide rates. Note that anxiety is the only psychological variable, and it is treated socially in terms of rate. This is a social explanation of social facts.

Of course, more data are needed before Durkheim’s theory can be fully accepted. (For example, can you think how one might try to measure group attachment or shared values?) One must also control for other relevant variables, such as age and marital status. These variables might affect the associations we have found. We know, for example, that single people have a higher suicide rate than do married people. Is it possible that Catholics are more likely to get married early and stay married longer than Protestants? If so, how might that affect the association between religion and suicide rates?

In order to state accurately that one variable causes another, we must have all three elements we have mentioned:

1. An association between the variables:
2. A time order, in which cause is prior to effect:
3. Away to control for other relevant variables.

Table analysis is one way of showing the first of these elements, and we have just discussed the importance of the other two. All causal analysis must begin with association. If there is no association, there can hardly be causality. You should now be able to discover associations in tables. (For more information on associations, causality, and controlling for the other relevant variables, see the sociology texts that have been recommended at the end of this section.)

Now you know of Durkheim’s methods. Try them in the following exercises.  

Exercise 1: Table Reading

Durkheim believed that the origins of homicide and suicide lie in different social conditions. Where the constraint of the moral order is weakest, isolation produces despair, depression, and suicide. Where it is strongest, homicidal retribution is permitted or even expected for violations of the moral order (which includes the ideas of personal or family honor and patriotic duty). Durkheim (1897/1951, p. 351) speculated that “where homicide is very common it confers a sort of immunity against suicide.” This explains the inverse relationship between homicide and suicide in Table 3. Table 6 uses information from the Statistical Abstract of the United States, 1994 (Table 133) to describe homicide and suicide statistics for a typical community of 1,000,000 people. Let’s use that information to investigate Durkheim’s idea.

Table 6: VIOLENT MALE DEATHS PER 100,000 BY RACE

<table>
<thead>
<tr>
<th>Causes of Violent Death</th>
<th>White</th>
<th>Black</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide</td>
<td>217</td>
<td>121</td>
<td>338</td>
</tr>
<tr>
<td>Homicide</td>
<td>93</td>
<td>720</td>
<td>813</td>
</tr>
<tr>
<td>Total</td>
<td>310</td>
<td>841</td>
<td>1151</td>
</tr>
</tbody>
</table>


---

5 MY NOTE: ENTER EXERCISE ANSWERS IN YOUR UNIT 2 PACK, PAGES 8 & 9.
6 MY NOTE: IN THIS COURSE THE INDEPENDENT VARIABLE WILL ALWAYS APPEAR ON THE COLUMN RATHER THAN THE ROW AS IT OCCURS IN THIS ILLUSTRATION.
(DUPLICATE QUESTIONS IN UNIT 2 PACK, PAGE 8)

1. How many variables are shown in this table? _________
2. How many values are there for each variable? ________
3. How many black males died violently?
4. How many white males died violently?
5. Did blacks or whites have the greater percentage of the total dying violently?
   Group _________
   Percent _________
6. How many men died from homicide? N = __________
7. What racial group had the greater percentage of homicides? Group _________ Percent _________
8. What racial group had the greater percentage of suicides? Group _________ Percent _________
9. Is there an association shown in this table? ______________ Explain your reasoning:
   ______________________________________
10. Which is the independent variable? ________________ Why?
11. Does homicide confer an immunity against (that is, reduce) suicide? Discuss the significance of these data.

Exercise 2: Multivariate Table Reading

As a final exercise in table reading, study the table below. With more than two variables, this table is more complex than the table in Exercise 1. Remember to follow the principles of table reading. Read the title and identify the variables in the table. Look in the cells and see if the data cluster on a diagonal. If there is no diagonal, is there another pattern that describes the relationships between the variables? Now answer these questions:

### Table 7: 1991 Suicide Rates by Sex, Race, and Age Group

(Rates are per 100,000 population. Excludes deaths of nonresidents of the United States.)

<table>
<thead>
<tr>
<th>Age</th>
<th>Total(a)</th>
<th>Male White</th>
<th>Male Black</th>
<th>Female White</th>
<th>Female Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Ages</td>
<td>12.2</td>
<td>21.7</td>
<td>12.1</td>
<td>5.2</td>
<td>1.9</td>
</tr>
<tr>
<td>10—14</td>
<td>1.5</td>
<td>2.4</td>
<td>2.0</td>
<td>0.8</td>
<td>(B)</td>
</tr>
<tr>
<td>15—19</td>
<td>11.0</td>
<td>19.1</td>
<td>12.2</td>
<td>4.2</td>
<td>(B)</td>
</tr>
<tr>
<td>20—24</td>
<td>14.9</td>
<td>26.5</td>
<td>20.7</td>
<td>4.3</td>
<td>1.8</td>
</tr>
<tr>
<td>25—34</td>
<td>15.2</td>
<td>26.1</td>
<td>21.1</td>
<td>5.6</td>
<td>3.3</td>
</tr>
<tr>
<td>35—44</td>
<td>14.7</td>
<td>24.7</td>
<td>15.2</td>
<td>7.2</td>
<td>2.9</td>
</tr>
<tr>
<td>45—54</td>
<td>15.5</td>
<td>25.3</td>
<td>14.3</td>
<td>8.3</td>
<td>3.0</td>
</tr>
<tr>
<td>55—64</td>
<td>15.4</td>
<td>26.8</td>
<td>13.0</td>
<td>7.1</td>
<td>2.1</td>
</tr>
<tr>
<td>65—74</td>
<td>16.9</td>
<td>32.6</td>
<td>13.8</td>
<td>6.4</td>
<td>2.4</td>
</tr>
<tr>
<td>75—84</td>
<td>23.5</td>
<td>56.1</td>
<td>21.6</td>
<td>6.0</td>
<td>(B)</td>
</tr>
<tr>
<td>85+</td>
<td>24.0</td>
<td>75.1</td>
<td>(B)</td>
<td>6.8</td>
<td>(B)</td>
</tr>
</tbody>
</table>

B) Base figure too small to meet statistical standards for reliability of a derived group.

Includes other races not shown separately.

(a) Includes other races not shown separately.
(b) Includes other age groups not shown separately.

(DUPLICATE QUESTIONS IN UNIT 2 PACK, PAGE 9)

1. How many variables are shown in this table? _________ Variable names: 
2. Identify the dependent variable. _______________________
3. Identify the independent variable(s). ____________
4. In the United States in 1991, how many suicides were there per 100.000 population? 
5. What group had the highest suicide rate? ______
6. What group had the lowest suicide rate? ____________
7. What relationship exists between age and suicide? 
8. What relationship exists between sex and suicide? ______
9. What is the most important relationship that exists between or among the variables? Explain how you reached your conclusion and what the association or lack of association means. ________________________________

PLEASE SEE ME FOR REFERENCE LIST.

STATISTICAL TOOLS
FOR DESCRIPTION AND COMPARISON

Society is not a mere sum of individuals. Rather, the system formed by their association represents a specific reality which has its own characteristics. The group thinks, feels, and acts quite differently from the way in which its members would were they isolated. If, then, we begin with the individual, we shall be able to understand nothing of what takes place in the group.

Emile Durkheim
*The Rules of Sociological Method*

Descriptive Statistics
Averages: Mode, Median, Mean
Comparisons: Percentages and Rates
Correlation: Scattergrams, Meaning, Significance, Rho

Averages and Comparisons

When asked to explain our behavior, our first response is to look inward and offer a set of unique individual motives. In fact, the cause of most of our social behavior is a reaction to forces external to us. To detect social forces and understand how they shape our behavior, we must take advantage of the tools that have been developed by statisticians. As we have seen, prior to Durkheim many believed suicide was a psychological fact. By looking at individual cases, they failed to see the social patterns that were distinctive of groups and societies. Their research should have begun with descriptive statistics, through which they could have summarized what were normal numbers (averages) of suicides in a society, made comparisons (through percents and rates) between societies and groups, and investigated whether social facts such as divorce and suicide were associated (correlated). In the following section, we will review three of the descriptive tools that are necessary to discover and understand the nature of social facts.

**Averages**

Just as it is often difficult to see the forest for the trees, it may be difficult to see the group for the individual. So many and so interesting are the differences that set individuals typical or normal. A statistical answer to this is the average. Averages are usually expressed as a mode, median, or mean. The mode is the most frequent observation. If there is only one mode, the distribution is unimodal. If two

---

7 MY NOTE: IN THIS COURSE, WE WILL RESERVE THE USE OF THE WORD ‘CORRELATION’ FOR ACTUAL STATISTICAL CORRELATIONS.
observations are equally frequent, the distribution is bimodal. If you have a bimodal distribution, check to see whether there are two distinct groups that have been combined. This might happen, for example, if males and females were aggregated into a common distribution of height. The result would be a bimodal distribution, composed of the modal height for males and the modal height for females.

The **median** is the middle number in an array of scores. An array refers to the arrangement of scores in ascending or descending order. The median is the middle location in the array of scores. It does not consider score values but score ranks. The median is valuable when trying to represent extremely uneven scores. For example, we may wish to describe the average income of people working in a small business. Assume their monthly income is …:

- $500
- 500
- 500
- 700
- 800
- 1,000
- 1,250
- 1,750

\[
\text{MODE} = 500
\]

\[
\text{MEDIAN} = \frac{(N+1)}{2} = \frac{10+1}{2} = \frac{11}{2} = 5.5 \quad \text{median}
\]

\[
\text{MEAN} = \frac{\text{Sum } X}{N} = \frac{17,500}{10} = 1,750 \quad \text{mean}
\]

Note that if the number of scores in an array is odd, the median will actually be the middle number of the array. If there is an even number in the array, the median will be between the two middle scores. In the illustration above, the two middle scores are $700 and $800. The midpoint between these two scores is $750.

The mean is the sum of the values of a set of scores, divided by the number of scores in that set. This is expressed by the formula

\[
\text{Mean } X = \frac{\text{Sum } X}{N}
\]

Where

- \( X \) = Score
- \( N \) = Number of scores

Since the mean uses the actual values of the scores, it is responsive to extreme values. Note in the illustration of monthly income the considerable difference between the mode ($500), median ($750), and mean ($1,750). They are all averages, but with different assumptions. The mode identifies the most frequently occurring score, the median identifies the middle-ranking score, and the mean identifies the average value of the scores. Which one value correctly describes an array may be as much a matter of perspective as a technical question. In labor relations, management would probably prefer the mean wage to bolster their claim of how good wages are, and the workers would prefer the median wage to support their claim of low wages. Thus, the facts do not “speak for themselves.” As producer and consumer of statistical evidence, you must actively involve yourself in the interpretation of statistical data.

**Comparisons**

**Percentages:** In our investigations we will need to compare groups. Are there differences in suicide rates between blacks and whites? Between men and women? Between rich and poor? If groups are of equivalent size, direct comparisons can be made (such as between 50 males and 50 females). But when there are great differences in size (such as between 87 blacks and 13 whites), direct comparison is difficult. Percentages standardize results of unequally sized groups by the frequency per 100 cases. This is expressed by the formula

\[
\% = \frac{f}{N} \times (100)
\]

Where

- \( f \) = Frequency
- \( N \) = Number of cases
Assuming 9 out of 20 students in a class were Catholic, for example, we would describe the class as 45 percent Catholic. Since the purpose of percentages is to simplify comparisons, round decimals off to the nearest tenth.

\[ \% = \frac{f}{N} \times 100 = \frac{9}{20} \times 100 = 45\% \]

To make accurate comparisons, it is important to know the numbers involved in computing the percentages. To help the reader interpret the meaning of a percentage table, report the total number of cases upon which the percentages are computed. If the number of cases is less than 30, it is preferable to simply report the number of cases and omit percentage comparisons. Unless the number of cases upon which the percentage is based is over 50, do not attach too much significance to small percentage differences.

**Rates:** When the number of occurrences of an event falls below 1 in 100, it is often easier to use a rate instead of a percent. In such cases, rates allow you to compare whole numbers instead of fractions of a percent. Rates are typically based on the number of occurrences per 1,000, 100,000, or 1,000,000 people.