

Thinking Through Statistics

Exploring Quantitative Sociology



James V. Spickard

with the assistance of

Janaki Spickard-Keeler

Toroverde Press
2005

THINKING THROUGH STATISTICS

Exploring Quantitative Sociology

with

Sociological Insights[®]

Software

*download software for free from
www.toroverde-press.com*

JAMES V. SPICKARD

with the assistance of

JANAKI SPICKARD-KEELER

Toroverde Press
2005

About the authors:

Jim Spickard teaches sociology and anthropology at the University of Redlands in Redlands, California, and formerly taught research methods at the Fielding Graduate Institute, Santa Barbara, California. He publishes widely on social theory, social research methods, and the sociology of religion. He developed the software and the approach used by this book over many years of university teaching.

Janaki Spickard-Keeler minored in mathematics at Smith College and until recently taught math and physics at Olney Friends School in Barnesville, Ohio. She now works in Washington D.C. She collected data for this project and helped develop several of the chapters.

TOROVERDE PRESS
30545 Bridlegate Drive
Bulverde, Texas 78163

Copyright © 2005 by James V. Spickard. All Rights Reserved

If this book did not come with a software CD, please download the Sociological Insights[®] software and data from:

- www.toroverde-press.com
- www.mcguire-spickard.com/software/software.htm

Downloads are free, as are updated datasets (which will be periodically available).

The authors wish to thank Ole Riis, Meredith McGuire, Judy Citron, and especially Dmitri Spickard-Keeler for their comments on our manuscript. Thanks to Gary Schulman and Rae Newton for noting statistical inconsistencies in a previous draft. An extra thanks to Dmitri for his suggestions about improving the Sociological Insights[®] software.

THINKING THROUGH STATISTICS

EXPLORING QUANTITATIVE SOCIOLOGY

TABLE OF CONTENTS

Quantitative Sociology: An Introduction	1
1. Mapping Social Regularities	9
• Aggregate Data	
2. Descriptive Statistics	19
• Health and Death Rates	
• Crime Rates	
3. Scatterplots and Correlations	31
• More Crime Rates	
4. Multiple Regression (I)	45
• Social Instability and Suicide	
• Catholics and Abortion; Baptists and Murder	
5. Multiple Regression (II)	57
• Religion, Social Instability, and Suicide	
6. Distributions and Cross-Tabulations	67
• Survey Data	
• Who's Happy?	
7. Indexes	77
• Who's Sexist?	
8. Control Variables	87
• Sex, Race, and Politics	
9. T-tests and Analysis of Variance	95
• Race, Education, and Socio-Economic Stratification	
• Divorce	
10. Standardized Cross-Tabulations	111
• Family Structure, Wealth, and Happiness	
Epilogue: Doing Sociology	123
Appendices:	
• Using the <u>Sociological Insights</u> [®] Software	131
• Variable Lists:	
• Data for the 50 U.S. States	135
• 2000 General Social Survey	141
• Index of Key Concepts	153

IMPORTANT!!

If this book did not come with a software CD, please download the Sociological Insights[®] software and data from:

- *www.toroverde-press.com*
- *www.mcguire-spickard.com/software/software.htm*

Downloads are free, as are updated datasets (which will be periodically available).

QUANTITATIVE SOCIOLOGY

AN INTRODUCTION

This book is designed to introduce you to quantitative sociology. “What’s that?” you say. Simply put:

Quantitative sociology is the study of the patterns underlying social life, using numbers.

These patterns are many. From the fact that property crime rates are higher in the western United States than they are in other parts of the country to the fact that lower-status parents expect more obedience from their children, there are many aspects of social life that only emerge by counting.

Quantitative sociology examines such things. It seeks to uncover hidden trends. Following a long sociological tradition, it often shows that “the emperor has no clothes” – by showing that what “everybody” thinks about social life is not true at all. Sometimes its revelations are surprising, and sometimes they are mundane. In either case, they tell us things about society that we might not otherwise know.

- Did you know, for example, that the abortion rate is higher in those U.S. states where a high proportion of the population is Catholic? It’s true, as we shall see in chapter 4. This is just the opposite of what we would expect, given the Catholic Church’s well-known opposition to abortion. But there is more to this figure than meets the eye.
- Did you know that states with high percentages of Baptists also have high murder rates? As we shall see, this does not mean that Baptists are doing the murders, but it may mean that Baptist sermons are less effective at setting a non-violent social tone than religious people think.
- We have all heard that those who marry young have a higher rate of divorce than those who wait a while before settling down. But what other factors influence the divorce rate? Among which social groups is divorce the most common?
- Does wealth bring happiness? I don’t mean, can it buy it. Are wealthy people happier than others, when one accounts for the fact that such people also tend to be well-educated – a group that tends to be happier than their less-educated peers?

We will explore these questions – and many others like them – in the pages that follow.

TWO KINDS OF DATA

This book enables you to do real social analysis using real social data. These data are of two kinds.

First, we have what we call **aggregate data** – data based on geographic regions. We could use any such regions, but here we will use the 50 United States. Aggregate data include such things as a region’s population and the percentage of that population that is female, tall, or unemployed. Such data include birth rates, crime rates, illness rates and anything else that happens to a population in proportion to its numbers. They also include more esoteric things: the percentage of the population with hunting licenses, the Cosmopolitan subscription rate, and the percentage of alcohol consumed as beer (as opposed to wine or hard liquor). Such data allow us to compare the social climates in various locales.

Aggregate data give us information about geographic regions

The aggregate data in this book come from a variety of sources. Information about the U.S. population at large comes from the Census Bureau. Crime figures come from the FBI Uniform Crime Reports. Health data come from the Department of Health and Human Services and from various private foundations. More esoteric figures come from such sources as Newsweek and USA Today. All of it is relatively accurate, though none of it is perfect. Not even the FBI gets everything right. But it is the best data we have.

Survey data result from questions asked of individuals – all of whom are members of some group or social category. They thus tell us about individuals – and by extension, about the groups themselves. The survey might cover a specific topic, such as political or religious attitudes, beliefs about sex education, and so on. Or it might be a more general poll that covers dozens of topics. The survey might have a small target population, such as dirt bikers or Congregationalists. Or it might cover the U.S. population as a whole. In each case, we not only learn about each individual’s opinions; we also learn how many individuals in that group have what kinds of opinions. How many Americans identify themselves as Catholics, for example? Do members of Generation X really have different political opinions than Baby Boomers? Well-constructed surveys can answer these questions.

Survey data give us information about individuals.

Obviously, we can’t ask such questions of the entire American population. We can’t even poll every Episcopalian or every skinhead. The solution is to ask a **sample** of our target population. We poll a small number of individuals, whose opinions represent those of people like them. We just have to choose that sample so that it represents the whole.

We won’t go into the mathematics of sampling in this book. All we need to know is that any sample has to be randomly selected, such that each individual in the target group has an equal chance of being chosen. And it needs to be large enough so that we can be reasonably confident that the range of views in the sample matches the range of views in the group itself. Fortunately, it turns out that a relatively small sample can accurately represent a very large population. Just over 1500 people, randomly selected, can model everyone living in the U.S.!

This book uses survey data from the 2000 General Social Survey (GSS). Conducted nearly annually from 1972 to 1992 and every two years thereafter, this is the most accurate regularly repeated survey of the U.S. population. Every even-numbered year, the National Opinion Research Center (NORC) interviews a random sample of about 3000 non-institutionalized American adults. They don’t cover people in jails, prisons, hospitals, or mental wards. But they do get a representative sample of pretty much everyone else who speaks English well enough to answer the survey’s 600 questions. This is a large enough sample to model the attitudes of the American public to within a few percentage points – excluding, of course, the jailed, hospitalized, etc. and those who don’t speak English. The GSS is the best survey in the business. Exploring it will show us what quantitative sociology can do.

ANOTHER WAY OF THINKING ABOUT DATA

There is another important way of thinking about data, one that adds some detail and helps us figure out what statistical techniques are most useful in any given situation. This is the threefold distinction among:

- **Categorical data** are data divided into discrete categories, like sex – “Men”/“Women” – race – “White”/“Black”/“Other” – or region – “Northeast”/“Midwest”/“South”/“West”. These divisions are common in survey data, where they are used to see how different types of people differ on various other factors. We might, for example want to know whether

people from the Northeast, Midwest, South, or West are happier, wealthier, etc. than people from other regions of the country. Categorical data let us investigate such things.

- **Ordinal data** are data that can be rank-ordered, but that do not lend themselves to exact measurement. For example, we might classify things as “Small”/“Medium”/“Large”/“Extra-Large”. These definitely have an order, but we have no idea whether the distance between “Small” and “Medium” is equal to the distance between “Large” and “Extra-Large”. These divisions are also common in survey data, where we ask people such questions as “In general, are you *very happy*, *pretty happy*, or *not too happy*?” We know that different people will have different definitions of what constitutes “very happy” versus “pretty happy”, but we also know that they will rank them in the same order. This ranking is all that matters for ordinal data.
- **Interval/ratio data** are data that can be both ranked and measured – exactly enough to be used in mathematical calculations. This includes such things as crime rates, disease rates, poverty rates, wealth ratios, population counts, and so on. For example, Louisiana’s murder rate of 10.7 per 100,000 residents is nearly 53% higher than Michigan’s rate of 7.0 per 100,000 residents. We can compare such things with ease.

Some survey data is also interval/ratio data. One example is a General Social Survey question that asks people how many years of schooling they have had. This lends itself to mathematical calculations rather easily, as we can tell whether one group of people has had half the schooling, twice the schooling, etc. as another.

(The GSS also asks for the highest degree that people have earned, which is an ordinal measure. “Not High-School”/“High-School”/“Jr College”/“College Degree”/“Graduate Degree” makes a lovely rank order, but it does not lend itself to exact calculation.)

Interval/ratio data are also sometimes known as **continuous data**, because there are no gaps between their possible values.

Under some circumstances, we can treat ordinal data as if they were interval/ratio data. We’ll see one such circumstance in Chapter 8. It pays to be wary, however, for such analyses can sometimes mislead.

Categorical data are made up of discrete categories: “Men”/“Women”, “White”/“Black”, etc.

Ordinal data can be rank-ordered but not measured with any precision: “Small”/“Medium”/“Large”, “Very Happy”/“Pretty Happy”/“Not Too Happy”.

Interval/ratio data can be both ranked and measured. They include crime rates, population counts, age in years (but not “Young”/“Middle-Aged”/“Old”).

KINDS OF ANALYSIS

We will encounter several kinds of statistical analysis in this book. They will be divided by chapter, as follows. The first five chapters use **aggregate data** from the 50 U.S. states.

- **Chapter 1** will introduce the concept of **distribution**. Using maps, we will see how different crime rates, population figures, and so on are distributed across the 50 United States. Where are crime rates high and where are they low? Where does a higher proportion of the population suffer from illness? Measuring distribution answers these kinds of questions.

- **Chapter 2** will introduce various **descriptive statistics**, among them measures of **central tendency** and **dispersion**. You have encountered these before, in your junior high school math class: an average is a measure of central tendency, while a range is a measure of dispersion. It makes a difference whether states' crime rates are all clustered around a single point, for example, or are spread out across a spectrum. These and other descriptive statistics will guide us here.
- **Chapter 3** focuses on ways of showing the relationships between two variables. **Scatter-plots** locate each of the 50 states on a grid, giving us a two-dimensional picture of their association. States with large urban populations, for example, also have high abortion rates – perhaps because there are more abortion clinics in cities. On the other hand, there is no particular connection between states with large numbers of old people and states with large numbers of college students. Some states are high on both, some low on both, and yet others are high on one and low on another.

This chapter will also cover **correlation** – a way to reduce the relationship between two variables to a single number. This is a particularly powerful tool.

- **Chapters 4 and 5** extend this discussion to the case of three or more variables. They cover **regression analysis** – a technique for sorting out the effect that one variable has on another, after eliminating the effect of a third, fourth, or fifth variable. To take an example I've already mentioned: abortion rates are higher in states with lots of Catholics and also in states with high urban populations. Do Catholics really have more abortions than members of other churches? Or do they just live in urban areas, where abortions are more readily available? Regression lets us decide.

Chapters 6 through 10 use **survey data** from the General Social Survey.

- **Chapter 6** explores the **distribution** and **cross-tabulation** of survey data. Distribution just means the number of people answering a question in a given way. For example: 75% of Americans would let an atheist give a speech in their community, while 24% would not; 43% think that the government should do something to equalize people's incomes, 35% think it should not, and 20% are neutral on the issue.*

Cross-tabulation, like correlation, involves discovering whether differences on one variable accompany differences on another. This gives us more detailed comparisons. To continue the previous example: 83% of religious liberals would allow an atheist to speak, but only 66% of religious fundamentalists would. That's a big difference! Similarly, 57% of people making less than \$15,000 per year think that the government should help equalize incomes, but only 29% of those making over \$110,000 a year agree. This is no surprise, but it is still a big difference – one that tells us a lot about American society.

- **Chapter 7** introduces the notion of indexes. An **index** is a single value that summarizes answers to a series of questions. For example, only 23% of Americans think that men are emotionally better suited than women for politics, and only 40% think that men should be the breadwinners while women manage the household. But 47% agree with at least one of these two statements – a better indication of sexism than is either question alone. Indexes often make it easier to uncover latent features of social life. There are several of them in our GSS survey data.
- **Chapter 8** expands cross-tabulations to include **control variables**. These essentially divide our sample into parts. Once divided, we run a separate cross-tabulation for each part. This helps us discover whether the social pattern that we see in one part of the population

* The percentages do not add to 100% due to rounding.

also occurs in another.

For example, American men and women are equally sexist: 48.8% of men and 46.9% of women give one or more sexist answers to the questions that make up the Sexism Index. (Such a small difference in the sample tells us that there is probably no difference in the total population.)

If we divide our sample into generations, however, we see that this equality only holds for men and women born before 1972. Women born after 1972 are significantly less sexist than are the men their age. Controlling for generation reveals a previously unnoticed pattern.

- **Chapter 9** demonstrates the value of **t-tests** and the **analysis of variance (ANOVA)**. These tests compare the average scores of different groups of people on one or another variable, to see whether they are different. (They also compare those people's dispersion around that average.) For example, people with college degrees have significantly higher incomes and occupational prestige, on average, than do people who only graduated from high school. This isn't true for everyone; many high school graduates have better-paying and higher-status jobs than do college folks. But the average college graduate does better on both measures than the average high school graduate does.

T-tests and ANOVAs especially help out when we don't have enough interview subjects to interpret a cross-tabulation. Despite the large number of people interviewed for the GSS each year, this happens more often than we might think. Using these routines with survey data is okay, so long as we are careful about which variables we use. It doesn't make sense to take an "average" of race or sex, for example. We have to watch what we are doing.

- **Chapter 10** introduces a tool with two uses. **Standardized cross-tabulations**, like t-tests and ANOVAs, give us another way of handling small numbers of interview subjects. And, like regressions, they let us measure the influence of several variables on another variable simultaneously. We know, for example, that people whose mother divorced are more likely to be divorced themselves. We also know that people who grew up in the Northeast are less likely to be divorced, and people who grew up in the West are less likely ever to have married. Which has more effect on one's marital status: one's parents' marital experience or one's region-of-origin? Standardized cross-tabulations help us answer such questions.
- Finally, the **Epilogue** suggests some sociological projects that these techniques can help you undertake. Pursue these, and you'll be doing real sociology!

NO MATH!

In the "bad old days," college undergraduates couldn't do these kinds of analyses. Not only did they call for great mathematical ability. They also required access to large computers and the ability to write complex software. It is one thing to calculate a correlation coefficient or a cross-tabular chi-square by hand. It is quite something else to solve a five-factor regression equation using matrix algebra, or generate a standardized cross-tabulation. Even my graduate statistics courses did not cover such complex topics.

That was unfortunate, for regression analysis and standardized cross-tabulations are powerful tools, even for undergraduates. Like some of the other tools we'll encounter in this book, they help us determine the true causes of social phenomena, which can sometimes be hard to see. Specifically, they help us separate the true causes of social phenomena from all of the false "causes" that might lead us to false conclusions.

The fact that I now carry a pocket computer with more processing power than the room-sized computers of my student days does more than just tell you my age! It also tells you how much more we can explore real sociology than we used to. The computer revolution has opened doors that we never thought possible. Sociology today is much more exciting, precisely because the less-mathematical of us can leave the math behind.

THINKING THROUGH STATISTICS

I have titled this book “Thinking Through Statistics” for a good reason. The kind of sociological thought that I am trying to encourage involves using statistical techniques to help us understand society better. The statistics themselves are just a means to an end. This is not a traditional statistics text! Unlike such texts, it does not concern itself with the mechanics of statistical analysis. It does ask you to learn to read various kinds of data tables – a task that should not prove too daunting. But mainly it focuses on setting up sociological research problems. It helps you learn how to:

- Think about what sociological questions you want to answer, and what data might best answer them;
- Form a hypothesis about what relationships between the data you expect to find;
- Pick the most appropriate test from among a limited set of statistical routines;
- Set up the relationships between your data so that you can test your hypothesis;
- Let the computer do the work;
- Read the results.

It’s that simple! Before you know it, you will be doing quantitative sociology.

The whole process is so simple, in fact, that I have often given problems like those in this book to introductory sociology students without mentioning the word “statistics”. On the last day of the semester, I let them know that they have been doing statistical analysis for four months. Most are not surprised, but some are. None are afraid of statistics ever again.

The math-phobes among you should probably ignore the word “statistics” whenever you run across it in the pages that follow. In fact, you will lose nothing by crossing it out – and by ignoring the footnotes, to which I have exiled my more overt statistical commentary. You can learn a lot by focusing entirely on the sociological conclusions that can be drawn from state-by-state and questionnaire data. What is more important than the details of how at this point.

Those of you who are more advanced will want to pay special attention to how to set up the various research problems that we encounter. You should focus on how to organize quantitative sociological knowledge and how to draw conclusions from the statistical results. You will probably want to read the footnotes, as these contain interesting details. But if you want to master pure statistical analysis, you should consult another text. Thinking through statistics, not getting stuck in them, is the agenda here.

SOCIOLOGICAL INSIGHTS

The key to this book is a computer program, Sociological Insights[®]. Depending on which edition of this book you have, you will either:

- find it on the enclosed CD, which will run on any Windows95 (or later) computer, OR
- download it from one of the web sites listed on the back of the book’s title page.

Designed specifically to highlight sociological thinking rather than mathematics, this program lets you analyze both aggregate and survey data without losing sight of the sociological purpose behind such analysis. The aggregate data comes from the 50 U.S. states, and the survey data comes from the 2000 General Social Survey. I have left a detailed description of this program to the chapters that follow, and a quick introduction – for those who need one – to the Appendix. For now, let me just say a bit about the program’s philosophy.

Most other programs for teaching quantitative sociology err in one of two ways. Either they are so complicated that students have to focus on learning the software rather than on putting that software to use. Or they are little more than slide shows, pushing canned presentations at students rather than letting them explore.

Sociological Insights[®] uses real social data, while remaining simple and transparent. You begin by following the examples in this book. You are not, however, limited to these examples. The program lets you explore the data in any way you want. In every one of my courses, several students discover social patterns that I have not considered before. This is not surprising, considering that the GSS dataset alone allows for over 15 million different controlled cross-tabulations!

Still, the most important part of this program is not its data, but the ease with which it lets users test sociological hypotheses. Sociological Insights[®] is powerful without being obtrusive. It helps you learn how to analyze sociological problems and, thus, how to think sociologically in a way that was previously not possible.

HOW TO READ THIS TEXT

Every book has its typographic conventions and this one is no different. Some of these involve core statistical concepts. Others involve my in-text instructions about how to use the Sociological Insights[®] software. I expect you to follow along with your software as you read, because this is the best way I know of to learn to think sociologically. I have tried to make this as easy as possible by taking advantage of the following printing customs:

CONCEPTS

- I use **boldface** when I introduce a core word or concept, and sometimes include it in a box alongside the text. I often underline the term when we encounter it again. I cease underlining when I have used the term enough to make its meaning clear.
- I include a list of “Things to Remember” at the end of each chapter, which includes that chapter’s core concepts.
- I have included an index at the end of the book so that you can find the definitions of core words easily.

SOFTWARE

- Each time we use the Sociological Insights[®] software, I provide instructions in **boldface** so that you can work through the examples as you read. Particularly toward the start of the book, I tell you which menu to use, which item to choose from that menu, and which variable number to type into the appropriate box. (Toward the end of the book, I sometimes skip some of these steps. By then, however, using the program will be second-nature.)

For example, on page 9 I tell you to:

- **Start Sociological Insights[®]**
- **Open the STates menu**
- **Click on “Map & List Values”**
- **Type “4” into the blank**
- **Click on the “OK” button**

These tasks require basic computer skills. You need to know how to start a program in Windows (double-click on it), how to use your mouse to open a menu (single-click on it), how to use your mouse to choose a menu item (single-click on it), how to type a number (without the quotation marks) into an onscreen box, how to use the mouse to press an onscreen button (single-click on it), and – ultimately – how to close a window (use your mouse to click the “Close” button or just hit the “ESC” key).

If you have any trouble figuring out how to do these things, I suggest that you read the instructions that came with your computer.

- Each variable has a number, which the software uses to identify it. I always tell you which variable number to use for our examples – usually in **boldface**, between quotation marks. Type in the number, not the quotation marks. (The program won’t let you type in quotation marks, but there is no point in having you get frustrated trying.)
- I almost always give you the variable name as well as the number. For example, on page 9 I write:

Type “4” (Var 4: %URBAN 00) in the blank.

This tells you to type the number 4 into the on-screen box (leave off the quote marks). You can read the variable name as “Percent Urban in 2000”. That stands for “the percentage of each state’s population that resided in metropolitan areas at the time of the 2000 U.S. Census”. You can find such expanded definitions in the list of variables – available at many places in the program – or by choosing the “Map & List Values” item from the **STates** menu and reading the expanded description at the bottom of the screen.

- If you have any trouble understanding such instructions, I suggest that you read the “Downloading and Using Sociological Insights[®]” appendix, starting on page A-1. The program is easy enough that even the most rabid technophobe can get the hang of it pretty quickly.

VARIABLE LISTS

- A list of the variables in the States dataset begins on page 131.
- A list of the variables in the 2000 GSS dataset begins on page 137.

ONE: MAPPING SOCIAL REGULARITIES

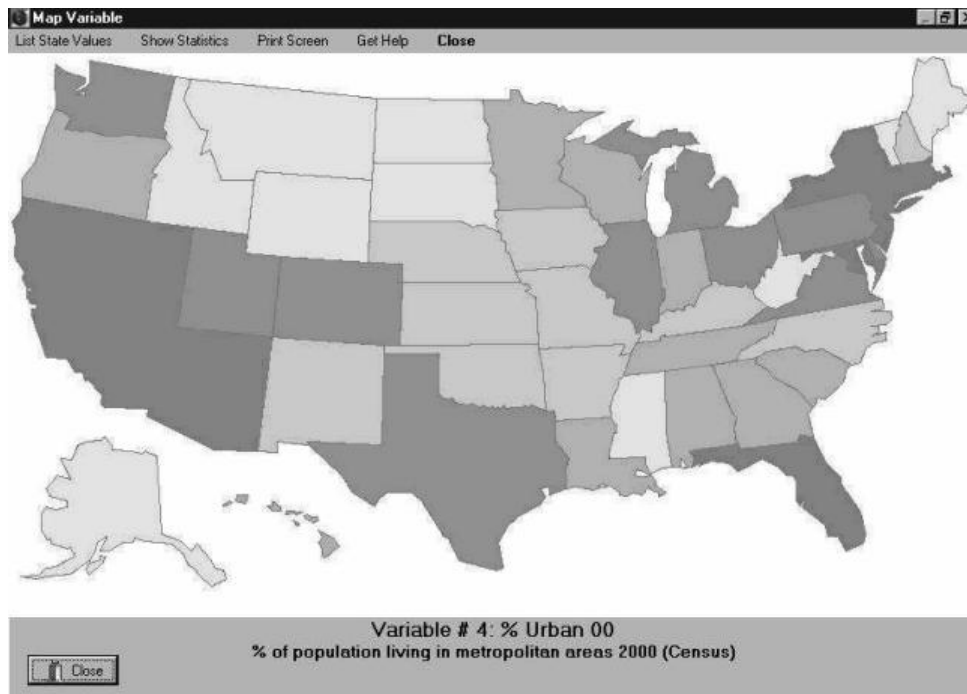
This chapter explores **social regularities** – one of sociology’s main topics. Regularities are stable patterns of social life. They come in several varieties, the most important of which are:

- Social patterns that last over time.
- Aspects of social life that seem to be joined, at least during a particular era.

Sociologists are interested in regularities because they are predictable. Once we know a pattern, we can do a much better job of solving social problems – one of sociology’s main products. But sociology also seeks to increase people’s self-awareness. Social patterns are usually unconscious – by which I mean that people don’t often see the larger patterns in which they take part. If people act in predictable ways, but do not know that they do so, then sociology can teach them something about themselves.

This chapter explores some of these regularities using **maps** that display **aggregate data**. Aggregate data are the numbers of things, people, or events found in a given area. How many people live in each of California’s counties? How many convenience stores are there in each of Chicago’s neighborhoods? How many copies of *Cosmopolitan* or *Playboy* are sold in each of the 50 U.S. states? These are all examples of aggregate data. They tell us something about social life in given regions. They also let us compare regions, to find out whether they are alike or different.

Let’s look at some aggregate data. **Start *Sociological Insights*[®]**,* then **open the *S*States menu** (in the upper left-hand corner of the program screen), and **click on “Map & List Values”**. **Type “4”** (Var 4: % URBAN 00) – without the quote marks – into the blank, then **click on the “OK” button**. You’ll get a screen like this:



* For more detailed guidance, see “Downloading and Using *Sociological Insights*[®]” starting on page A-1.

This map represents the percent of each state’s population living in urban areas, as measured by the 2000 U.S. Census. The 10 states in dark red are the most highly urbanized; the 10 states in light pink are the most rural, and the rest are arrayed (in groups of 10) in between.

Aggregate data give us information about geographic regions

New York, Connecticut, and New Jersey shouldn’t surprise us, because everyone expects them to be urban. But Texas? Isn’t that the land of wide open spaces? It is – but those spaces are wide open because nobody lives in them. Over 80% of Texans live in Houston, Dallas, San Antonio, Fort Worth, Austin, and other big cities.

Find and press the “List State Values” menu button in the upper left-hand corner of the screen. A box will appear, containing a list of states **ranked** from most to least urban. Here are the top and bottom thirteen:

Ran	State	Value
1	NEW JERSEY	100.00
2	CALIFORNIA	96.70
3	MASSACHUSETTS	95.90
4	CONNECTICUT	95.60
5	RHODE ISLAND	94.10
6	FLORIDA	92.80
7	MARYLAND	92.70
8	NEW YORK	92.10
9	ARIZONA	88.20
10	NEVADA	87.50
11	ILLINOIS	84.90
12	TEXAS	84.80
13	PENNSYLVANIA	84.60

Ran	State	Value
38	ARKANSAS	49.40
39	KENTUCKY	48.80
40	IOWA	45.30
41	NORTH DAKOTA	44.20
42	WEST VIRGINIA	42.30
43	ALASKA	41.50
44	IDAHO	39.30
45	MAINE	36.60
46	MISSISSIPPI	36.00
47	SOUTH DAKOTA	34.60
48	MONTANA	33.90
49	WYOMING	30.00
50	VERMONT	27.80

Of the top ones, I find Arizona and Texas the most surprising. But that’s because I’ve spent a lot of time on their Interstate Highways, which are pretty deserted. The bottom ones don’t surprise me, though I would have guessed that Alaska was a bit lower than it is. Like Texas, Alaska has a lot of empty space, but 41% of Alaskans live in the cities.

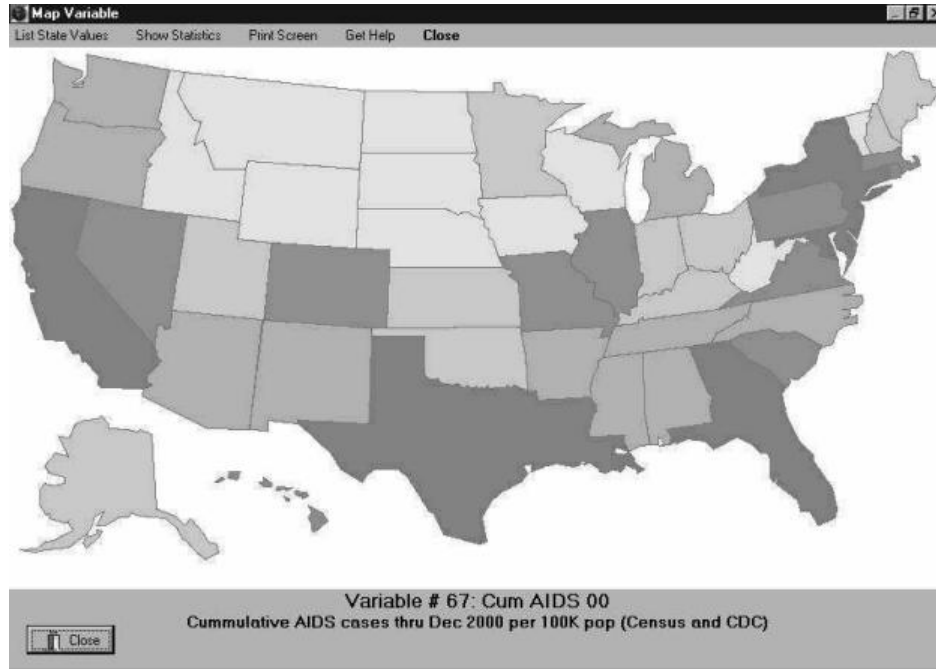
Like ordinal data, **interval/ratio data** are rank-ordered. However, the distance between the rankings can be measured accurately. This makes possible mathematical calculations that are not available for use with other data types.

Remember that we are dealing with rates here, not with absolute numbers. Rhode Island has only about 1/3 as many people as Iowa, but the percentage of its population that lives in urban areas is twice as large.

These are, by the way, **interval/ratio** data (as explained in the Introduction). We know that Pennsylvania is twice as urbanized as West Virginia, because the percentage of its population that lives in cities is twice as high. This lets us calculate exact relationships between states that merely dividing them into categories or rank-ordering them would not. The statistics that we will be meeting in this and the next four chapters only work with such **continuous** data.

POSITIVE CORRELATIONS

Close the list by **clicking on the “X”** in the upper right-hand corner of the list window. Take a good look at the map again, then **click “Close”** or **hit the <ESC> key** to close it. Now **type “67”** (Var 67: CUM AIDS) in the blank and **press the “OK” button**. You’ll see a map of the cumulative number of AIDS cases per 100,000 population. This is the total number of AIDS cases that occurred in each state from the late 1970s (when the epidemic began) through the end of December, 2000, divided by the state’s population.

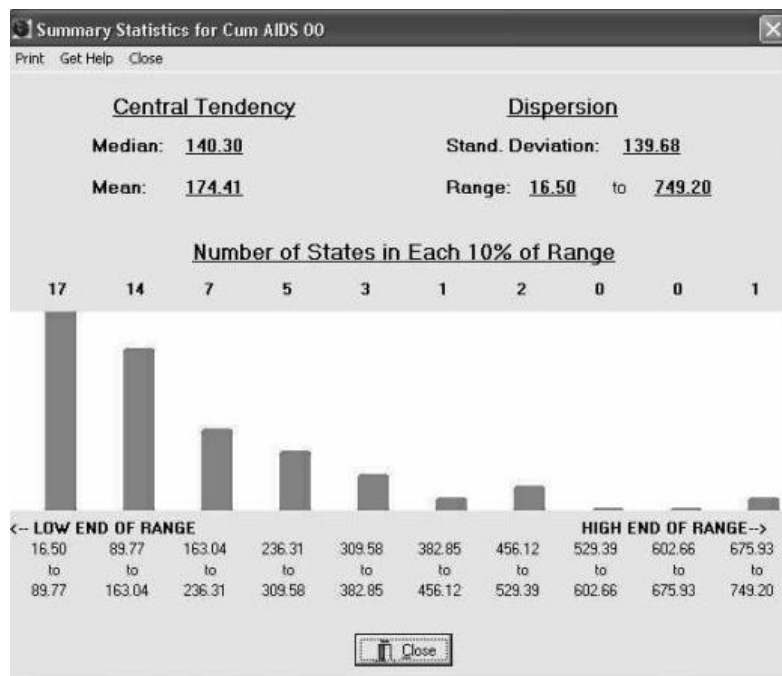


This map looks almost like the previous one! New York, New Jersey, California, Florida and Texas are high, states in the northern plains and Rocky Mountains are low, and the Midwest and South are in the middle. This looks just like the percent of the population that lives in cities. What might be going on?

If we think about it for a minute, this isn’t so strange. AIDS is largely an urban disease, transmitted by unsafe sex and by shared hypodermic needles. Cities have a higher amount of hard drug use, as well as a higher amount of sex between strangers. Moreover, epidemics have a hard time getting a toehold in rural areas, because the population is dispersed. It’s hard for those germs to find a new host in the wide-open spaces.

Again, **press the “List State Values”** menu item. New York, Florida, and New Jersey head the list, while North Dakota, South Dakota, and Montana are far behind. Figures range from 749.2 per 100,000 to 16.5 per 100,000. That’s a huge spread.

Close the List window by clicking the “x”, then **press the “Show Statistics”** menu button. You’ll see the diagram at the top of the next page:



For now, let’s ignore the figures and look at the graph. There are ten bars, each of which covers 1/10 of the range from 16.5 to 749.2 – the top and bottom values. Each bar shows how many states have AIDS rates in that portion of the range. The leftmost bar shows that 17 states fall between 16.5 and 89.8 cases per 100,000 population, the second leftmost shows that 14 states fall between 89.8 and 163.1 cases, the next shows that 7 states fall between 163.1 and 236.4 cases, and so on. This tells us that most states have relatively low levels of AIDS. Though the full range is from 16.5 to 749.2 AIDS cases per 100,000 population, the range in all but the top four states is about half that.*

There is an important lesson here about **distribution**. The maps divided the 50 states into 5 groups of 10 states each. They showed the highest ten in dark red, the next ten in semi-dark red, the next ten in medium red, the fourth ten in light red, and the bottom ten in pink. For AIDS, this put New York (749.20) in the same group as Texas (258.9). That’s a big difference! The maps tell us how the states rank relative to each other – in groups of ten. But they don’t tell us much about the actual AIDS rate in any given state. We need a distribution graph for that. A distribution graph tells us that a very few states have had a much larger AIDS problem than do most of the others.

We will spend more time with distribution graphs in the next chapter.

* * * *

What’s the point here? The two maps we’ve looked at so far tell us something about the relationship between AIDS and urbanization. They tell us that AIDS infection is more common in urban areas; urban areas are more likely to have high rates of AIDS. The two things – AIDS and urbanization – happen in the same states. When two variables go together like this, we say that they are **positively correlated**. This means that their maps look pretty much alike. As a general

* Remember that we are talking here about rates, not the absolute number of people living in urban areas or people with AIDS. Rhode Island has relatively few people, but 94.1% of them live in urban areas and 195.9 per 100,000 population have gotten AIDS. Michigan has a much larger population, and therefore more people with AIDS and more people living in urban areas. But the rate of both urbanism and AIDS is lower there. Don’t ever confuse rates of anything with the actual number of people involved.

trend, where one is high, so is the other; where one is low, so is the other. In this case, where there is more urbanism, the AIDS rate is higher, and where that rate is lower, there is less urbanism. The two things seem to happen – or not happen – in the same places.

So: the maps look alike, and the rates of urbanism and AIDS are positively correlated. There is an important qualifier to this, however: Correlation does not necessarily imply cause. It may just so happen that AIDS and urbanism go together. Two social facts may just happen to accompany one another, without either one of them causing the other. For example, AIDS certainly does not cause urbanization! That would be ludicrous. Any sociologist who made such a claim would become a laughingstock.

A positive correlation means that both things happen in the same places.

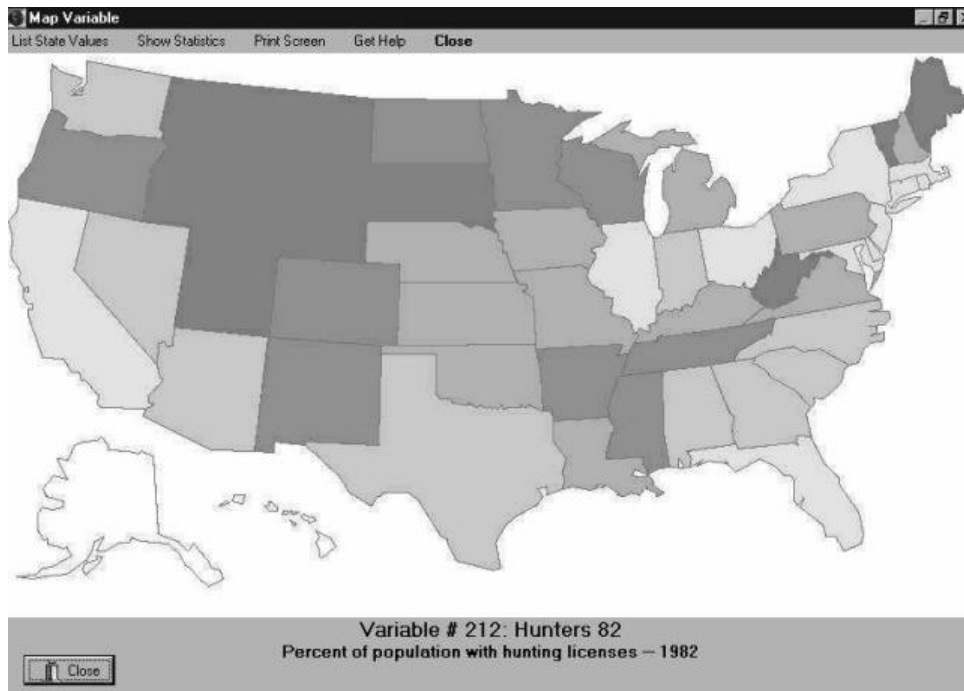
In this case, however, there is at least an implied causality running the other direction. Urbanism does not exactly “cause” AIDS, but concentrations of people in urban areas make it much easier for a disease like AIDS to spread. Thus, a high degree of urbanism contributes to the AIDS rate, without causing the disease directly. Having lots of people concentrated in a small area makes an AIDS epidemic possible.

Not all of the positive correlations that we will encounter in this book are causal, even to this limited degree. Part of the fun of quantitative sociology is figuring out which ones are causal and which are not.

In any event, these two maps tell us that urbanism and the AIDS rate are **positively correlated**. If we know that one of these figures is high for a particular state, we can predict that the other will be as well. If we know that one of them is low, we can predict the same of the other. That is what “positive correlation” means.

NEGATIVE CORRELATIONS

Close the AIDS map, then type “212” (Var 212: HUNTERS 82) in the blank. This shows us a map of the percent of the population having hunting licenses in 1982. Yes, these figures are old. But they don’t vary much, even from decade to decade – and we don’t have recent figures.



Take a close look at this map, then compare it to the other two maps in this chapter. What do you notice about it? Clearly, it isn't like the others. In fact, it's pretty much opposite. Wyoming, Montana, Idaho, South Dakota, Vermont, Maine – these states top this list. They were at the bottom of the others. The big urban states – New Jersey, California, New York – are at the bottom of this one. They were at the top of the others.

(Alaska and Hawaii are blank because we don't have any data for them. **Anytime we don't have data for a state, the map displays it in white.**)

We call this situation a **negative correlation**. This simply means that where one thing is common, the other is rare, and vice versa. Where lots of folks have hunting licenses, few have AIDS. Where lots of people have AIDS, few have hunting licenses. Not only do these things not vary together – they vary oppositely! AIDS and urbanism happen in one kind of place; hunting happens in another. (Thinking of a photographic negative may help you remember this term.)

If we know that two things are negatively correlated, and we know that a particular state is high on one of them, then we can predict that it will be low on the other. This is the essence of negative correlation.

This is actually a very good example to remind us that correlation does not equal cause. This is true of negative correlations just as it is of positive ones. There is no direct connection between the AIDS rate and the rate at which the residents of a given state buy hunting licenses. True, people suffering from AIDS may be too sick to hunt, but the disease is relatively rare – it is certainly not common enough to depress the rate at which people purchase hunting licenses in New York City.

While there is no causal relationship between the AIDS rate and the purchase of hunting licenses, there is a likely causal relationship between urbanism and hunting. It's pretty hard to hunt in cities! I suppose that one could hunt at the zoo, or bang away at stray cats. But zoo-hunting will land you in jail and killing cats doesn't require a hunting license. It makes sense that states with a high percentage of people in cities would have a lower percentage of the population with hunting licenses. There is causality here; we just had to figure out what kind.

AIDS happens in cities, and hunting doesn't happen in cities. The **positive correlation** between AIDS and city life is real. The **negative correlation** between hunting and city life is real. But the negative correlation between AIDS and hunting is not real. It results from their independent connection with city life – positive for AIDS, negative for hunting.

We'll revisit cases like this in chapter 4. For now, you just need to remember three things:

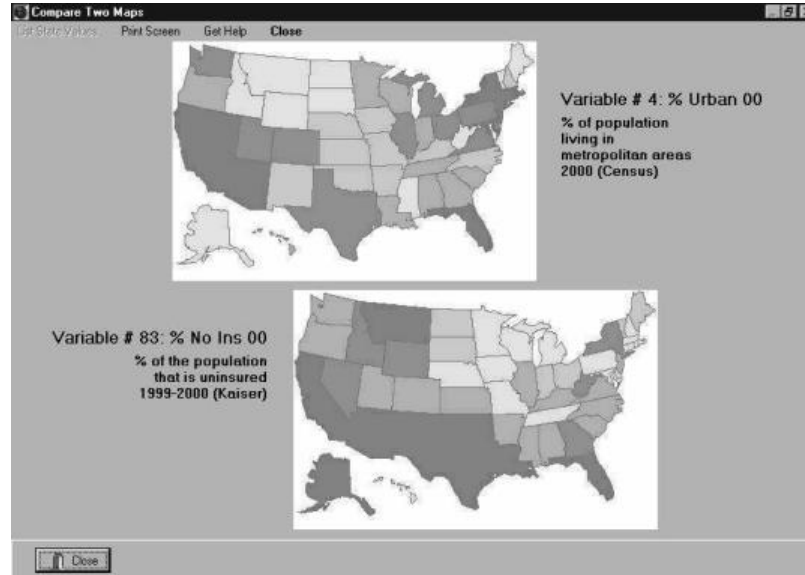
- Correlation is not (necessarily) cause.
- A **positive correlation** is where two things always happen in the same places.
- A **negative correlation** is when two things always happen in different places: where one is, the other isn't (and vice versa).

BEING UNCORRELATED

There is a third option, besides positive and negative correlations. Two social facts may be **uncorrelated**. This just means that there is no direct relationship between them. If a positive correlation means that two things happen together, and a negative correlation means that they happen oppositely, then a **lack of correlation** means that they don't happen either together or oppositely!

<p>A negative correlation means that two social facts are related, but opposite. Where one happens, the other doesn't.</p>

An example will help make this clear. **Close** the map of hunting licenses, and **hit the <ESC> key or the “Cancel” button** to return to **Sociological Insights®** main program screen. **Click on the “States” menu, then choose the “Compare Maps” option.** **Type “4”** (Var 4: %URBAN 00) in the first box and **“83”** (Var 45: NO INS 00) in the second one. Leave the other two empty **Click on “OK”**:



We know the first map. It shows the percentage of each state’s population living in urban areas. The second map shows the percentage of each state’s population that is not covered by health insurance. Values range from a high of 24% in New Mexico to a low of 6% in Rhode Island. Southern states (plus Montana) are generally high, while northern states are generally low.

Compare these maps. Do they look at all alike to you? Clearly not. Do they look like opposites? Again, no. To be the opposite of urbanism, the no-insurance map would have to look like the hunting-license map, which it does not. We know that urbanism and hunting are opposites: any map that is negatively related to urbanism would have to be positively related to hunting.

These two maps don’t look anything like each other, either positively or negatively. We call such relationships **uncorrelated**. There is no pattern to the relationship between uncorrelated variables.

When two things are **uncorrelated**, it means that there is **no direct relationship** between them.

Sometimes students confuse a negative correlation with two variables that are uncorrelated. The first indicates a relationship of opposites, like a photographic negative. The second indicates no direct relationship at all.

How about causality and a lack of correlation? This is a bit easier to grasp. If there is no connection between two things, then it is impossible for one of them to cause the other! A lack of correlation between two social facts means that there is no causal connection between them.

There is no direct relationship between urbanism and the lack of health insurance. If we have data on one of rates for a particular state, we can’t predict the other.

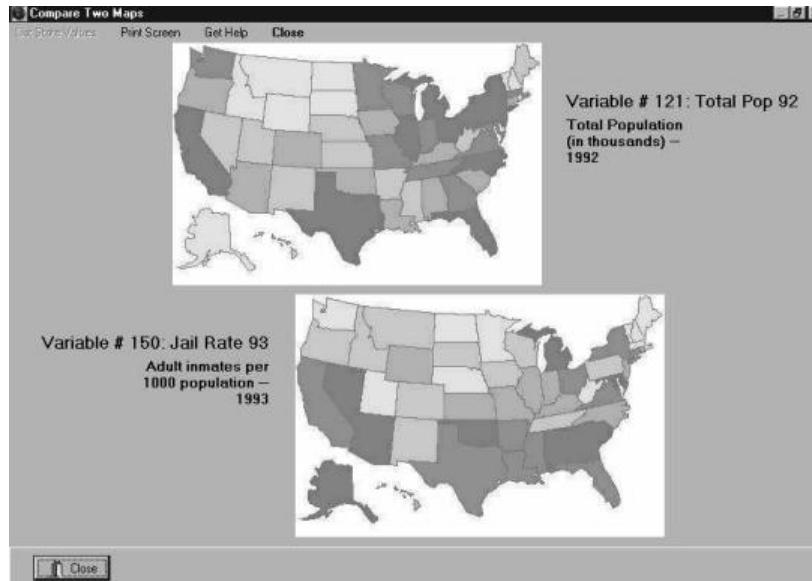
RAW FIGURES vs. RATES

So far, we have been mapping **rates**. Such things as the percentage of a state's population living in cities, the number of AIDS cases per 100,000 population, the percentage of a state's population owning a hunting license, etc. all describe a **number per unit of population**. Why do we do this? Aren't raw figures good enough?

Sociologists focus on rates because they help us make better comparisons between one geographic unit and another. Were we to use raw AIDS figures, for example, we would expect to find more AIDS cases where there are more people. This wouldn't tell us much – because we already know that California, New York, Texas, and Florida have the largest populations. It is much more useful to know how many AIDS cases there are relative to the total population.

Similarly, we would expect to find more people without health insurance in the states with the largest population – simply because there are more people there! But this would not tell us that New York does a much better job of providing health insurance for its citizens than do the other large states. Only rates tell us this.

Close the comparison between urbanism and health insurance, and **type "121"** (Var 121: TOT POP 92) and **"150"** (Var 150: JAIL RATE 93) in the first two boxes. Leave the last two boxes empty. **Click "OK"** to bring up a comparison between a map that shows a state's total population in 1992 and one that shows its jail rate a year later: the number of adult prison inmates per 1000 population in 1993. *



These maps are not very much alike. Nor are they particularly opposite. In fact, they are uncorrelated, though not so clearly uncorrelated as are the two maps that we previously viewed. California, New York, Texas, Florida, and Pennsylvania top the population list; Delaware, Nevada, South

* Though we have more recent figures for total population – Variable 3: TOT POP 00 – our most recent rate of incarceration is for 1993. We can compare data from 1993 with data from 1992 or 1994 without problems, because rates don't change much from year to year. They may, however, change significantly over ten years. We want to compare data taken at roughly the same time, so we use the older population figures rather than the latest ones.

Carolina, Alaska, and Alabama have the most citizens incarcerated per 1000 residents. North Dakota is near the bottom of both lists.

If the “Jail Rate” map contained raw figures, however, it would look much more like the “Population” map – just because the big states have much larger prison systems! The entire population of Alaska, for example, is smaller than California’s prison population. Were we to map raw figures, we would never know that Alaska locks up such a high percentage of its people.

As a general rule, sociologists compare rates rather than raw figures. And if we have only raw figures at hand, we make rates out of them. The census data in the STATES data set, for example, all began as raw figures: the number of single females or married couples in each state, etc. We turn these into rates by dividing by the population. We then report the result either as a percentage of the population or as a number per 1,000 or 100,000 population.

Some data don’t need this transformation. The FBI Universal Crime Reports already calculate crime rates for us, though they also report raw figures. One always has to be aware of the data one uses.

THINGS TO REMEMBER

- A **positive correlation** means that the maps are pretty much alike. Where one thing happens, so does the other. Where one thing doesn’t happen, the other doesn’t either.
- A **negative correlation** means that the maps are pretty much opposite. (Think of a photographic negative.) Where one thing happens a lot, the other doesn’t happen much – and vice versa.
- **No correlation** (or **uncorrelated**) means that the maps bear little resemblance to each other, either positively or negatively. There is no direct relationship between them.
- Sociologists generally compare **rates** rather than **raw figures**. We find these rates by dividing the raw figure for a state by that state’s total population. Comparing rates lets us identify the social differences between the various states and parts of the country – differences that have little to do with whether states are large or small.

TWO: DESCRIPTIVE STATISTICS

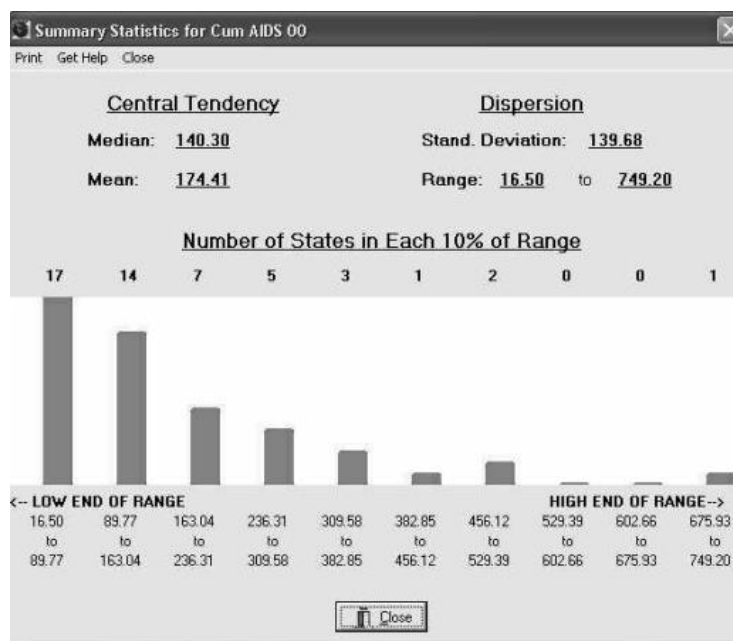
The last chapter introduced you to **aggregate data** – data based on geographic areas.* We looked at this data as it was displayed on a map of the 50 U.S. states. We saw the rate of urbanization, of AIDS infection, of hunting licenses, and so on. The exercises that accompanied that chapter introduced us to crime rates, census figures, and various American social problems.

These maps are all built in a similar way. The ten states scoring highest on a particular variable were displayed in bright red, the next ten in medium red, the next ten in a lighter red, and so on. This is one way of displaying that variable's **distribution**.

Descriptive statistics simply describe the data.

In the midst of that chapter, however, I showed you a different kind of distribution graph – of the cumulative rate of AIDS infection. (This was the total number of AIDS cases in each state from the late 1970s through 2000.)

Start **Sociological Insights**[®], then click the **States** menu, choose “**Map & List Values**”, and type “**67**” (Var 67: CUM AIDS) in the blank. Click “**OK**” to see the map. Then click “**Show Statistics**”.



As you remember, the AIDS rate ranges from 16.5 per 100,000 population (in North Dakota) to 749.2 per 100,000 population (in New York). **The graph divides this range into ten equal sections. The height of each red bar represents the number of states in each section.** The three sections at the high end of the range (on the right) have only one state between them. The three at the low end of the range (on the left) have 17, 14, and 7, respectively.

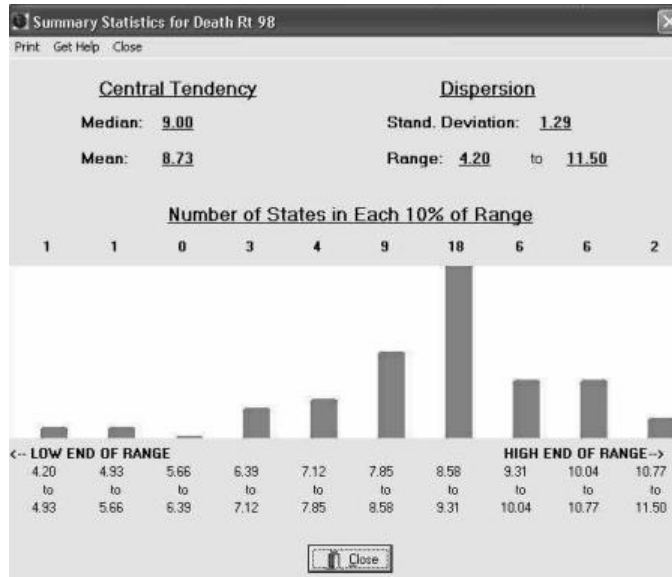
The graph and numbers on this screen describe the data. That's why we call them **descriptive statistics** – the topic of this chapter.

* Aggregate data can also be based on groups. We can, for example, compare poverty rates among Anglo- and African-Americans. The logic of such data is the same as with the geographic units that **Sociological Insights**[®] features.

Skew tells which side of the graph has the bulk of the values.

This graph has what statisticians call a **positive skew**. That just means that the bulk of the cases are toward the low end of the graph, while a few cases tail off towards the high end. A **negative skew** is just the opposite: where the bulk of cases are toward the high end and the tail is to the left.

The 1998 death rate is negatively skewed, as the graph at the top of the next page shows. Follow the same instructions, but **type “59”** (Var. 59: DEATH RT 98) to map the death rate. Then **click “Show Statistics”** to see the distribution.



Unlike the AIDS rate, the death rate does not vary much from state to state. It ranges from 4.2 per thousand in Alaska to 11.5 per thousand in West Virginia. Alaska is likely low because no one retires there. West Virginia and the other states at the top end of the range have large elderly populations. The bulk of the states fall just to the right of the center, with very few at the lower (left) end of the range. The negative skew is pretty easy to see.

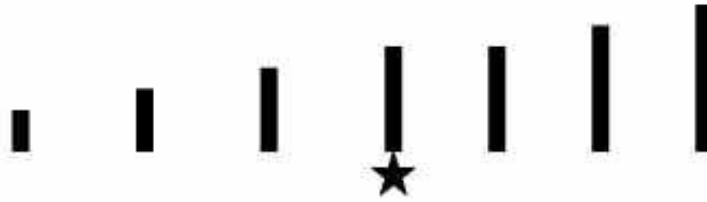
Look at the numbers above each of these graphs. They show the **median**, the **mean**, the **range**, and the **standard deviation**. These are what we call **descriptive statistics** because they *describe* our data.

- The **median** and **mean** are known as **measures of central tendency**, because they tell us something about the center of each variable’s distribution.
- The **range** and **standard deviation** are known as **measures of dispersion**, because they tell us something about how the various states’ values spread around this center.

Let’s first consider the measures of central tendency. What is a **median**? Imagine a series of lines of varying height:



If we reorder them by height, with the shortest on the left and the tallest on the right, we can identify the median as the one in the middle.



I have marked the median with a star. Exactly half of the values are above it and half are below it. This is simple with an odd number of values, but it's not much more difficult with an even number of values. You just count down from each end until you reach the two middle ones; the median is the average of these two.



“Median income” is the income right in the middle. Half the households earn more, half earn less.

In the two charts we've seen so far, the median cumulative AIDS rate is 140.3 per 100,000 population, which is much closer to the low than to the high end of the range. The 1998 median death rate is 9.0 per 1000, closer to the high end than to the low. Each of these is typical of **skewed distributions**. Negatively skewed distributions have medians closer to the high end of the range. Positively skewed distributions have medians closer to the low end.

Something else is typical of skewed distributions: the relationship between the **median** and the **mean**. You're all familiar with the **mean**, although you may not know it. The **mean** is the same as an **average**. To find it, you add up all the values and divide by the number of values – in this case by 50 (unless you're working with a variable that's missing some data). Adding up the AIDS rates for all the states gives us 8720.5; dividing by 50 gives us a mean of 174.41. Adding up all the death rates gives us 436.5; dividing by 50 gives us a mean of 8.73.

Notice that the AIDS rate median is 140.3 – lower than the mean of 174.41. Medians lower than means is typical of positively skewed distributions. The opposite is true of the death rate: the median is higher than the mean. This is typical of negatively skewed distributions.

Both **mean** and **median** tell us a lot about our data.

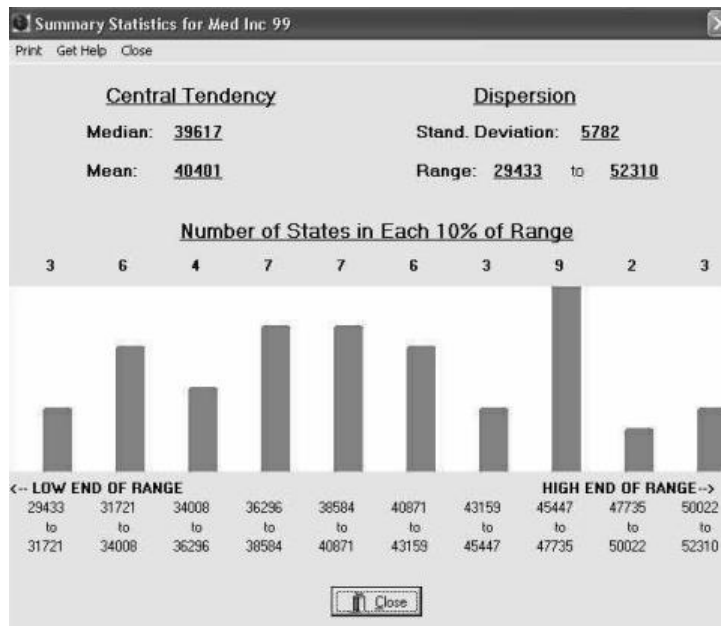
The **mean** gives us an easily calculable average figure, which makes for easy comparison. For example: the mean burglary rate in 1999 (Variable 105) was 720.2 burglaries per 100,000 population. The mean burglary rate in 1960 (Variable 228) was 402.25 per 100,000 – only 56% as high. Those 40 years saw nearly a doubling of the burglary rate – clearly a social trend.

Mean = average
Median = middle value

The **median** is often more useful when working with very skewed distributions. The burglary rate for 1960 is a good example, because just four states pulled up the mean significantly. We call these **outliers**, and they can really make a difference to our analysis. For example, the median burglary rate in 1995 was over 2.5 times the rate in 1960. Those four high states made the

difference in means lower. Comparing the means makes us underestimate the rise in burglaries. In this case, the median is a better tool for comparison.

Let's look at a less skewed distribution. Hit <ESC> twice to get back to the "Enter Variable to Map" window, type "34" (Var 34: MED INC 99), then click "OK". This brings up a map showing the median household income for each state in 1999. Click "Show Statistics" to show the distribution.



This distribution is not even, but it is also not particularly skewed. Maryland, Alaska, and Connecticut are at the top end of the range (the rightmost bar). West Virginia, Arkansas, and Montana are at the low end of the range (the leftmost one). The median is \$39,617, exactly what we find for Ohio. Twenty-five states lie above the median and twenty-four lie below it. This is a pretty even distribution.

This screen shows two **measures of dispersion** as well as the two **measures of central tendency**. One of these is the **range**. This is just the distance from the highest value to the lowest value. Maryland has the highest value at \$52,310. West Virginia has the lowest at \$29,433 – a bit more than half as much. The range is thus \$22,877, though we usually just list the lowest and highest figures. In this case, we would say: "In 1999, Median household income for the 50 states ranged from \$29,433 to \$52,310."^{*}

Ranges are pretty simple, but they can be misleading. Imagine a situation in which you have ten people earning the following hourly wages:

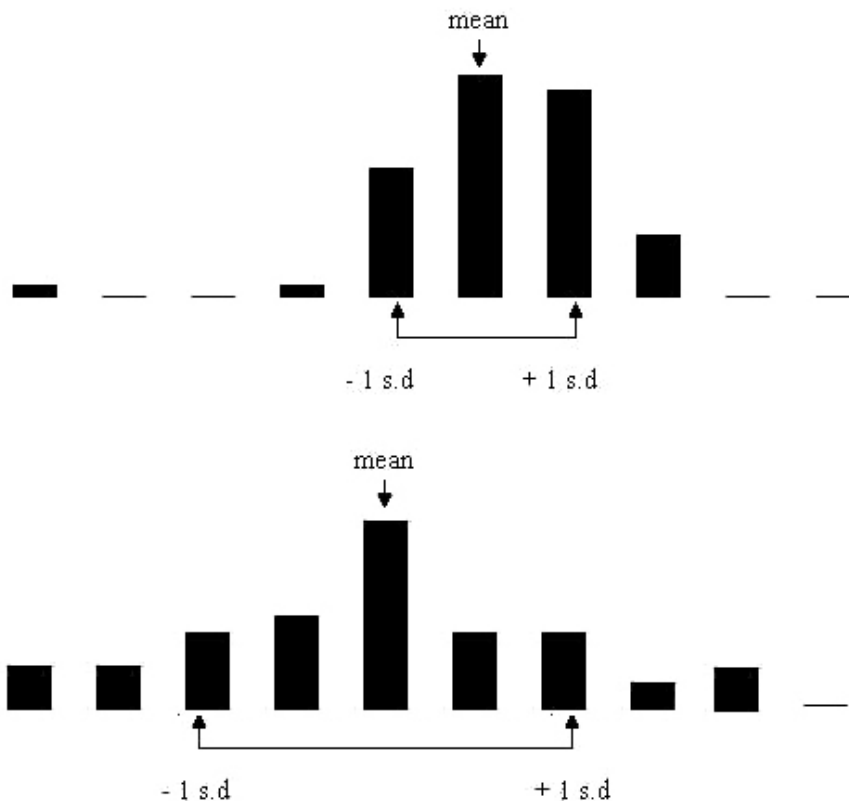
\$5 \$5 \$5 \$5 \$5 \$5 \$5 \$5 \$5 \$100

^{*} The U.S. Census Bureau often reports median income as a three-year moving average – i.e., as the average of the last three years of figures. This evens out the variation in year to year income levels. This gives us a more stable measure, because it lessens the chance that one state or another will be having an atypical year. Ideally, this gives us a more accurate estimate which states are consistently on top, which are consistently on the bottom, and which are consistently in the middle. For an example, check out Variable 138: Med \$ 90-92.

One can say, “These hourly wages range from \$5 to \$100.” This statement is accurate – but does it really describe the data? That one person earning \$100 per hour really distorts things. Is there a better way to talk about the spread of figures without being so completely pulled off track?

Sociologists have developed various ways to handle this. The most common of these is a statistic called the **standard deviation**. The math is tedious, but not terribly complex.* To oversimplify, the standard deviation amounts to the average distance of the various states’ values and the mean value. This gives you a number that describes how closely values cluster around the mean.

Sometimes values cluster rather tightly, as in this graph. Here, the mean is 6 and the standard deviation (s.d) is 1. On a ten-point scale, that’s pretty tightly clustered.



Still other times values are pretty spread out, as in this one. Here, the mean is 5 and the standard deviation is 2. That’s twice as large, which tells us that the values are more widely distributed.

In fact, so long as the distribution is not too skewed (as it is not in either of these cases), the standard deviation tells us about how many cases are within certain key distances from the mean:

- About 68% of the cases fall within one standard deviation of the mean, either above it or below it.

* Calculate how far each state is from the mean value – its deviation. Square each of these values. Add all the squared number together, then divide by the number of states. This gives you the average of the squared deviations – also known as the **variance**. Take the square root of that value. This is the **standard deviation**.

Descriptive Statistics

- Almost 95% of the cases fall within two standard deviations of the mean, either above it or below it.
- Over 99% of the cases fall within three standard deviations of the mean, either above it or below it.

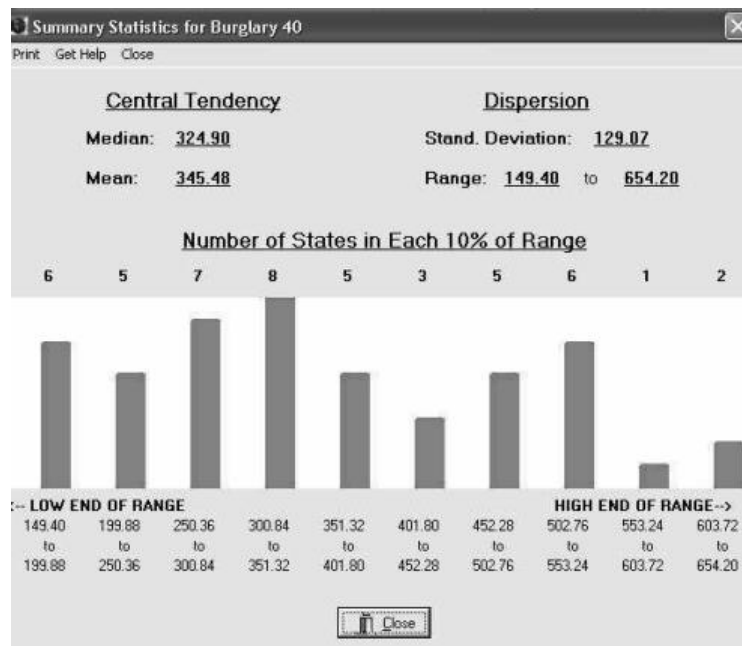
You may not know it, but you have all worked with standard deviations before. Both IQ tests and the Scholastic Aptitude Test (SAT) are scored in standard deviations. The mean IQ on both the Wechsler and Stanford-Binet tests is 100. The standard deviation on the Wechsler test is 15, so 68% of W-IQs are between 85 and 115. The standard deviation on the Stanford-Binet test is 16, so 68% of S/B-IQs are between 84 and 116.

Small standard deviations indicate tightly clustered data; large ones indicate that the data are spread out.

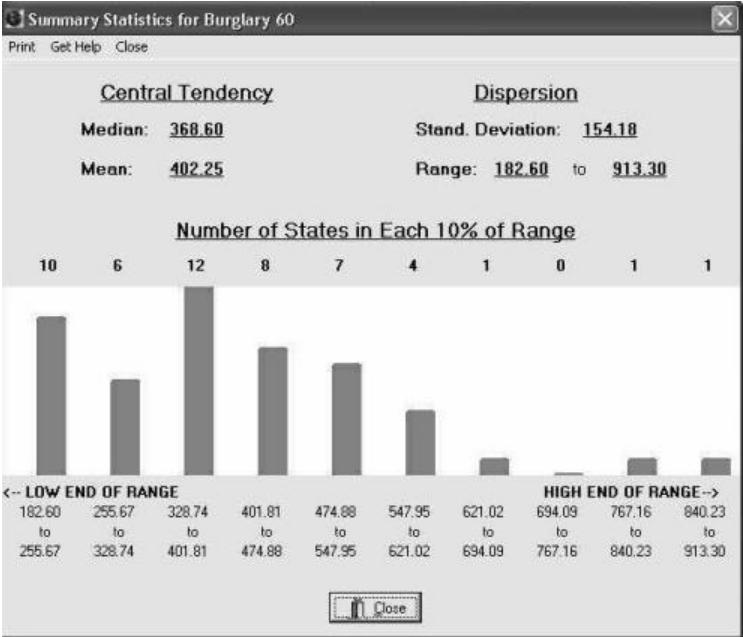
The mean score for each component of the SAT – both math and verbal – was originally 500 with a standard deviation of 100. This put 68% of all SAT results between 400 and 600 and 95% between 300 and 700. Note that I said “originally”. Over the years, Educational Testing Service has adjusted the SAT to make the results comparable from year to year. This involves shifting both the mean and the standard deviation each time the test is administered. I don’t know where they stand now.

A SOCIOLOGICAL COMPARISON

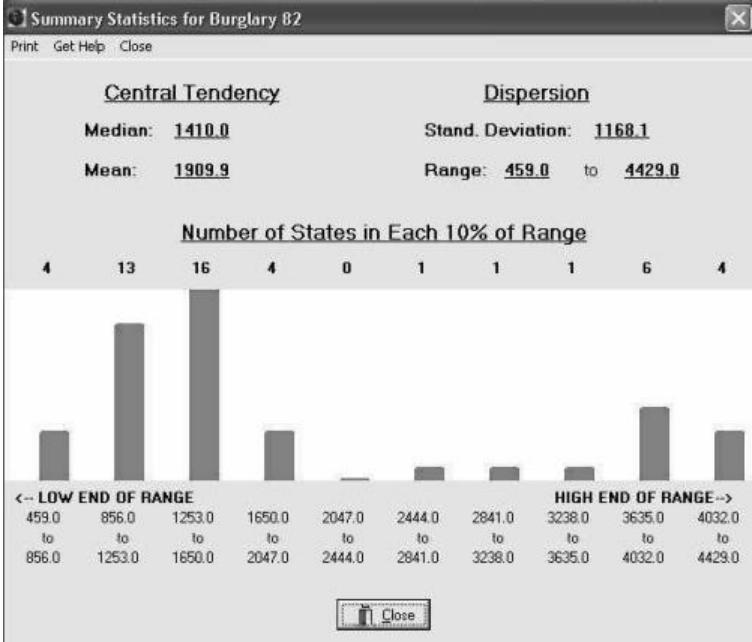
Let’s use these measures to make comparisons. **Hit <ESC> twice** to get back to the “Enter Variable to Map” window, **type “229”** (Var 229: BURGLARY 40), then **click “OK”**. This brings up a map showing the state-by-state burglary rate for 1940. **Click “Show Statistics”** to see the descriptive statistics.



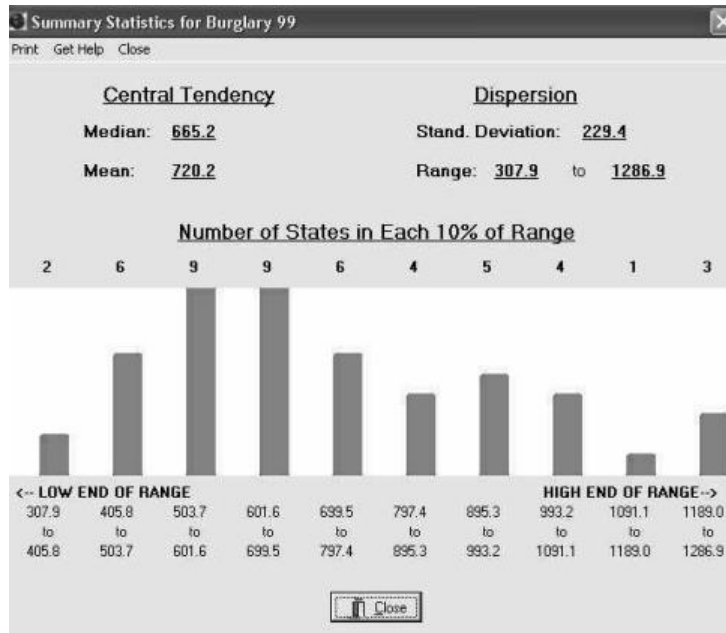
Now do the same for Var 228: BURGLARY 60 (the state-by-state burglary rate for 1960):



And Var 227: BURGLARY 82 (for 1982):



Finally: Var 105: BURGLARY 99 (the state-by-state burglary rate for 1999):



Compare these screens. What can we say about changes in the burglary rate over this 55 year span?

First, we'll look at the **measures of central tendency** – the means and medians:

	mean	median
1940	345	325
1960	402	369
1982	1910	1410
1999	720	665

Clearly, burglary rates have varied over time. Not only were they much lower in 1940 and 1960 than in recent years; they shot way up in the early 1980s. In fact, 1982 was the year of the worst economic recession in the U.S since the 1930s. Perhaps that is why burglary was so high then.

But check out the differences between means and medians. In 1940, the means and medians are pretty close to one another, which indicates a relatively even distribution. The mean and median were somewhat farther apart in 1960 and in 1999. But the 1982 mean is 500 points higher than the 1982 median. What's going on?

Remember that the mean is easily thrown off by a few outlying numbers. Check out the 1982 graph again, and you'll see that it is ridiculously spread. There's a lot of states toward the left side of the graph, whose burglary rates are not much higher than those of the other years. Then there are a few states way off to the right, whose burglary rates are astronomical. This is called a **bimodal distribution**, because the states cluster in two lumps rather than centering themselves on one. The graphs for 1940, 1960, and 1999 are not exactly **unimodal** (with one clear lump). But they are nowhere near as extreme.

Now let's look at **measures of dispersion** – the range and standard deviation:

	range	stand dev
1940	149-654	404
1960	183-913	492
1982	459-4429	3731
1999	307-1286	726

Clearly, 1982 is an odd year. Not only was that year's range wildly larger than were the ranges both before and after. But the standard deviation is huge. Indeed, the distance from one standard deviation below the mean to one standard deviation above it extends far past the actual lowest and highest values! Is this possible?

Actually, it is – and it makes hash of the idea that about 68% of a variable's values always lie within one standard deviation from the mean. **This works for unskewed distributions, but not for skewed distributions.** A standard deviation larger than the total range is a mathematical artifact – not something that you should worry about. It just underscores the degree to which burglary rates that year were not clustered around a central point.

You may be wondering why we bother with these numbers when we can just look at the graphs. Why trouble ourselves calculating anything beyond ranges when the graphs themselves tell us how different or alike one year is to another? The answer is that graphs and ranges work when we're dealing with a relatively small number of cases. Fifty cases is small for studying aggregate data. Sociological research typically uses much larger data sets, such as the 3000+ U.S. counties. Graphs don't work so well for these larger sets – and the graphs end up being less bizarrely shaped, as well. In these cases, **means, medians, ranges, and standard deviations** let us describe distributions rather well. They take up a lot fewer pages than graphs, besides.

What do these figures tell us about changes in burglary rates? Quite a bit, actually. Here's a part of the story:

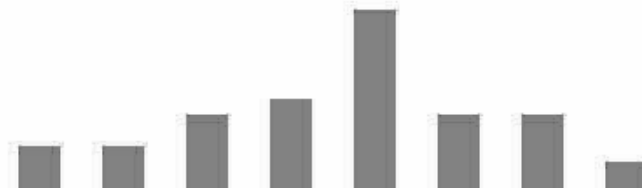
- First, we see that there has not been a steady rise in that rate from 1940 through today. Instead, the rate rose slightly from 1940 to 1960, spiked in the early 1980s, and then fell back to an intermediate figure.
- Second, we see that much of that 1980s spike occurred in a few states whose burglary rates went through the roof. (We know that from the bulge at the high end of the graph.) The burglary rate for most states was a bit higher than before, but not anywhere near the rate of those unfortunate few at the top of the scale.
- The drop from 1982 to 1999 was also not uniform. To some extent, it merely brought the states on the high end more in line with the majority.

The overall story is one of a differentiation between states, not of a steady rise in burglary for them all.

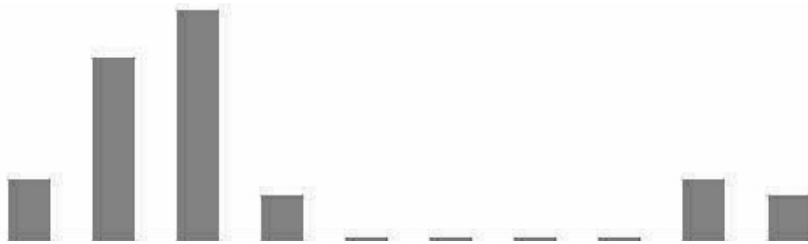
MODALITY

We need to visit one more measure of central tendency before leaving descriptive statistics. This one, however, is not so much mathematical as visual. It is the **mode**.

I've already used the terms **unimodal** and **bimodal** to refer to distributions. Basically we ask, "One lump or two?" The graph of the 1982 burglary rate is bimodal. The graph of the 1992 death rate shown early in this chapter was unimodal. A **mode** refers to the location of a graph's highest point: the point at which the largest number of cases clusters. The graph at the top of the next page shows a **unimodal distribution**, with the mode slightly to the right of center – slightly negatively skewed, to use the terminology introduced earlier. Here is a unimodal distribution (with a slight negative skew):



The next distribution is **bimodal** – with one mode to the left of center and the other far to the right:



Knowing a variable's **mode** tells us a lot about it. In the simplest cases, the mode, mean, and median all fall in the same spot. This tells us that the variable's values all cluster around one center. In more complex cases, such clustering does not occur. Bimodality is not rare – but it is particularly important.

Whenever we encounter a bimodal distribution, we want to know why some states cluster in one spot on the graph and others cluster elsewhere. There may be different social forces operating on each cluster. Once we notice the split, we can begin to explore the reasons for it.

Modes refer to the biggest lumps in a graph – where the most values cluster.

THINGS TO REMEMBER

1. The **median** and **mean** are known as **measures of central tendency**, because they tell us something about the center of each variable's distribution.
 - The **median** is the case in the middle – the one with an equal number of cases below and above it.
 - The **mean** is the average – the sum of all the values divided by the number of cases.
2. The **range** and **standard deviation** are known as **measures of dispersion**, because they tell us something about how individual states' values spread around this center.
 - The **range** is the distance from the lowest figure to the highest.
 - The **standard deviation** is a measure of how closely clustered values are around the mean.
3. **Measures of central tendency** and **measures of dispersion** are good for comparing one variable with another.
4. Distributions can be **positively skewed** or **negatively skewed**.
 - **Positive** skewness locates the bulk of the cases to the low end of the graph, with a few cases tailing off toward the high end. The mean is higher than the median.
 - **Negative** skewness locates the bulk of the cases to the high end of the graph, with a few cases tailing off toward the low end. The mean is lower than the median.
 - An **unskewed** distribution arranges the bulk of the cases in the middle of the graph, with smaller numbers of cases tailing off evenly to each side. The mean is close to or the same as the median.
5. Distributions can be **unimodal**, **bimodal**, **trimodal**, and so on. **Modes** refer to the points on the graph where there are the largest number of cases. Means, medians, and standard deviations are usually more informative about unimodal distributions. Bimodal, trimodal, etc. distributions indicate that different cases react differently to social phenomena.
6. In a reasonably large population with a unimodal, unskewed distribution
 - About 68% of the values fall within 1 standard deviation of the mean, either below or above it.
 - Almost 95% of the cases fall within 2 standard deviations of the mean, either below or above it.
 - Over 99% of the cases fall within 3 standard deviations of the mean, either below or above it.

THREE: SCATTERPLOTS AND CORRELATIONS

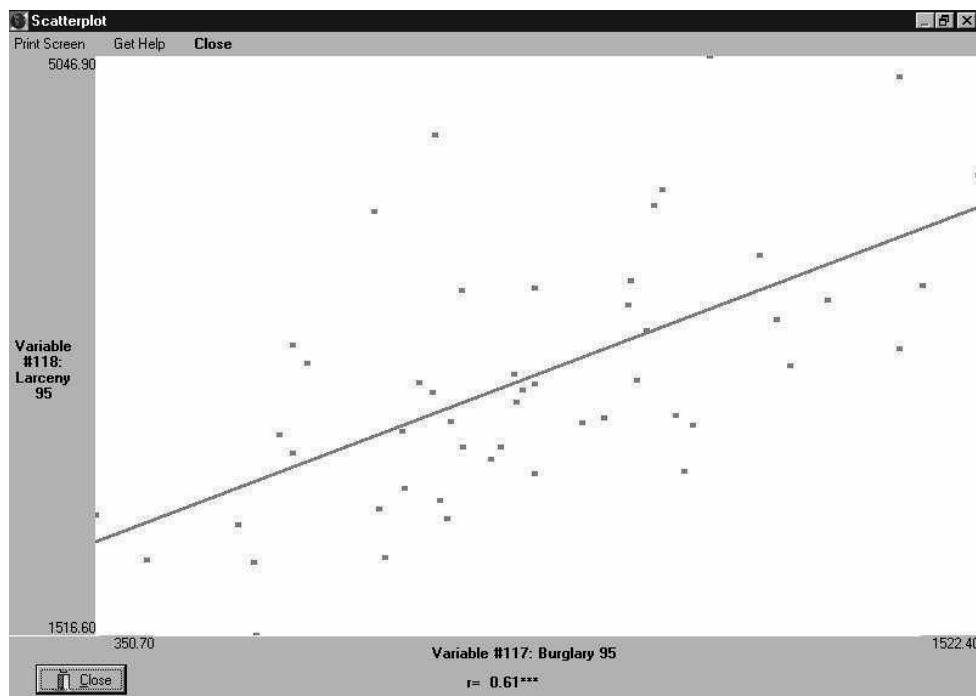
As you saw in Chapter 1, trying to tell by sight whether two maps are alike two maps is a bit difficult. Often, it seems impossible. Some maps look a lot alike, others look somewhat alike, and still others look a bit alike, but also a bit different. But one person's "somewhat" alike may be another's "a lot" alike. Eyeballing them isn't enough. We need a more precise way of measuring the differences between maps (and thus between the population levels, crime rates, etc. that they depict).

Fortunately, we have such a technique. It's called the **Pearson correlation coefficient**, and it was invented by Karl Pearson over 100 years ago. It gives us a number to represent the **positive correlations**, the **negative correlations**, and the **lack of correlations** that we explored in Chapter 1. This number is also sometimes called the **regression coefficient**, and is symbolized by an "r". (It is also called "**Pearson's r**", after its inventor.) When you see an "r" on your screen, that's the Pearson coefficient, and it tells you how alike two maps are.

Before we can understand correlation coefficients, however, we have to learn a bit more about what it means for two variables to "co-relate". We'll start with **scatterplots**, because they give us a good visual image of the extent to which different variables match one another.

SCATTERPLOTS

Start **Sociological Insights**[®], then open the "STates" menu, and click on "Scatterplot." Type "117" (Var 117: BURGLARY 95) in the first blank and type "118" (Var 118: LARCENY 95) in the second. Then click the "OK" button. You'll see the following graph:



At the bottom of the graph, you see the label "Variable #117: Burglary 95". That's the rate at which people's property is stolen when they are away. (Property stolen in their presence is called "robbery".) The burglary rates of the 50 states in 1995 are plotted along the X-axis of the graph: the number of burglaries per 100,000 population. They range from 351 (North Dakota) to 1522

(Florida). Florida residents are about 4 1/2 times as likely to have their homes burgled as are the residents of North Dakota. (But wouldn't you still rather live in Florida?)

At the left, you see the label "Variable #118: Larceny 95". "What's larceny?" you say. Larceny is financial theft and fraud: writing hot checks, using bum credit cards, forging bank drafts, etc. The 1995 larceny rates of the 50 states are plotted along the Y-axis of the graph.. These are per 100,000 population and range from 1517 (West Virginia) to 5047 (Hawaii). Larcenies are over 3 times as common in Hawaii as in West Virginia, relative to population size. Again, most people would rather live in Hawaii than in West Virginia, even if it means getting a few hot checks every now and then.

This X-Y graph works just like the graphs you studied in junior high school. We mark off a state's position on the horizontal (X) axis, and draw a vertical line running through it. Then we mark that state's position on the vertical (Y) axis, and draw a horizontal line through it. Where those two lines meet, we put a dot. We then repeat for each of the other forty-nine states.

If you want to try doing this by hand, get yourself some graph paper. Go back to the **Map & List** routine and copy down the values for each of the states for both Burglary and Larceny. Now find the two values for each state, tracking one on the horizontal axis and the other on the vertical. Pretty soon, you'll have a scatterplot just like the one on the previous page.

(Having fun yet? That's why we use a computer!)

Now look at the slanted line running from the lower left to the upper right corner of the scatterplot. This is called the **regression line**. It represents the straight line that comes the closest to connecting all the dots together. Obviously, no straight line could connect all those dots. But the regression line is the closest (mathematically) to doing so. This line highlights something important about the dots. See how the line slopes upward from right to left? Now see how the pattern of dots also slopes upward from right to left. States with high rates of burglary also tend to have high rates of larceny. States with low rates of larceny tend to have low rates of burglary. The dots show us that these two crimes – burglary and larceny – tend to happen in the same places.

Whenever the regression line slopes upwards from left to right, it indicates a **positive correlation**. As you remember from Chapter 1, two rates are positively correlated when they are both high in some places and both low in others. That's true of burglary and larceny. There are, of course, exceptions: Utah is #3 in larceny but only number #34 in burglary. But the two rates mostly vary together.

Look again at the bottom of the screen, just below the name of the X-axis variable. You see an "**r= 0.61*****". This is the **Pearson's "r"** that I mentioned on page 31. It is a measure of the degree of correlation between the X-axis and the Y-axis variables.

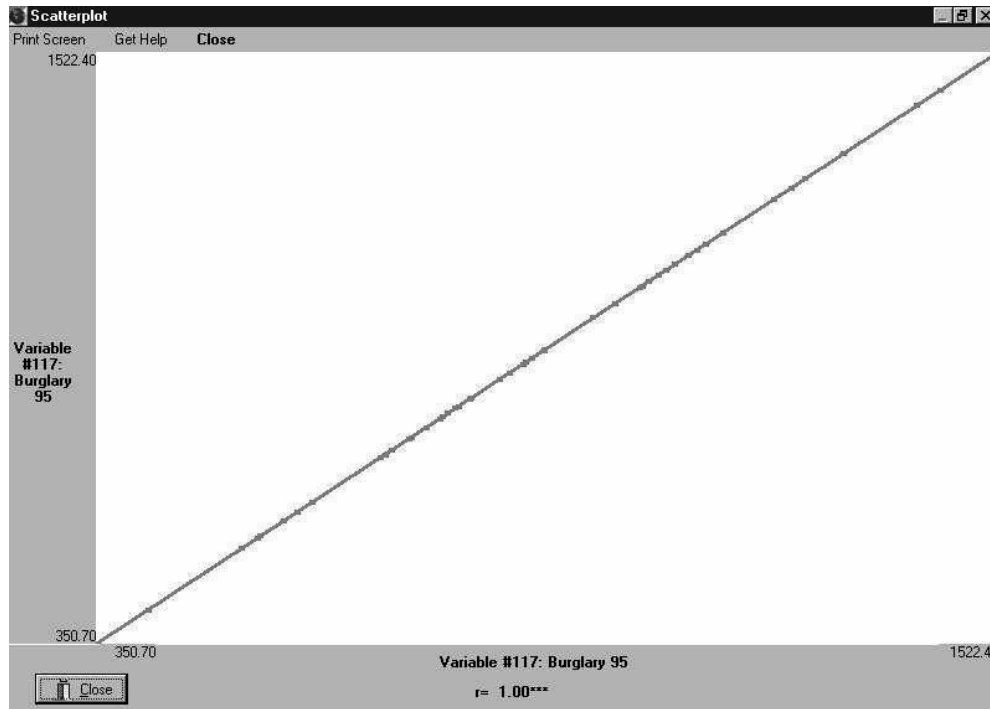
Both sign and size matter for the **Pearson's "r"**. A **positive** number indicates a **positive correlation**. A **negative** number indicates a **negative correlation**. A number **close to zero** indicates **no correlation**. The closer the number is to 1 or to -1, the stronger the relationship is between the variables. The closer the number is to 0, the weaker the relationship.*

Let's look at the case of a **perfect positive correlation**. Close the

When you see an "**r**" on your screen, that's the **Pearson correlation coefficient**. It tells you how alike two variables are.

* Mathematically inclined students will want to know that the Pearson coefficient measures the degree to which the dots cluster around the regression line. Conceptually, it is based on the sum of the vertical distances between the dots and the line. An **r** close to 0 means that the dots don't cluster at all; an **r** close to 1 (or to -1) means that the dots cluster tightly along the line. See page 36.

scatterplot, then **type “117”** (Var 117: BURGLARY 95) for BOTH the X-axis and Y-axis variable. Here’s the graph:



What's the most striking thing about this graph? Quite obviously, all the dots are right on the regression line. This makes sense, because every dot has the same value on both X- and Y-axes. We should expect all the dots to line up on a diagonal -- and that's just what they do.

Now take a look at the "r=" at the bottom center of the screen. It says "**r=1.00*****". This is as high as **Pearson's r** can ever get. It means that the two variables are perfectly correlated. The highest state on one variable is the highest state on the other; the second on one variable is the second on the other; and so on, all the way to the lowest. A perfect correlation usually means that the two variables are measuring the same thing – and we certainly are in this case.

One seldom gets a perfect correlation in social science, but the "**r=0.61*****" correlation between burglary and larceny we saw in the earlier graph is pretty close. The burglary and larceny rates in the 50 states are very strongly correlated with one another (which, as you remember, means that the two crimes happen in the same places).

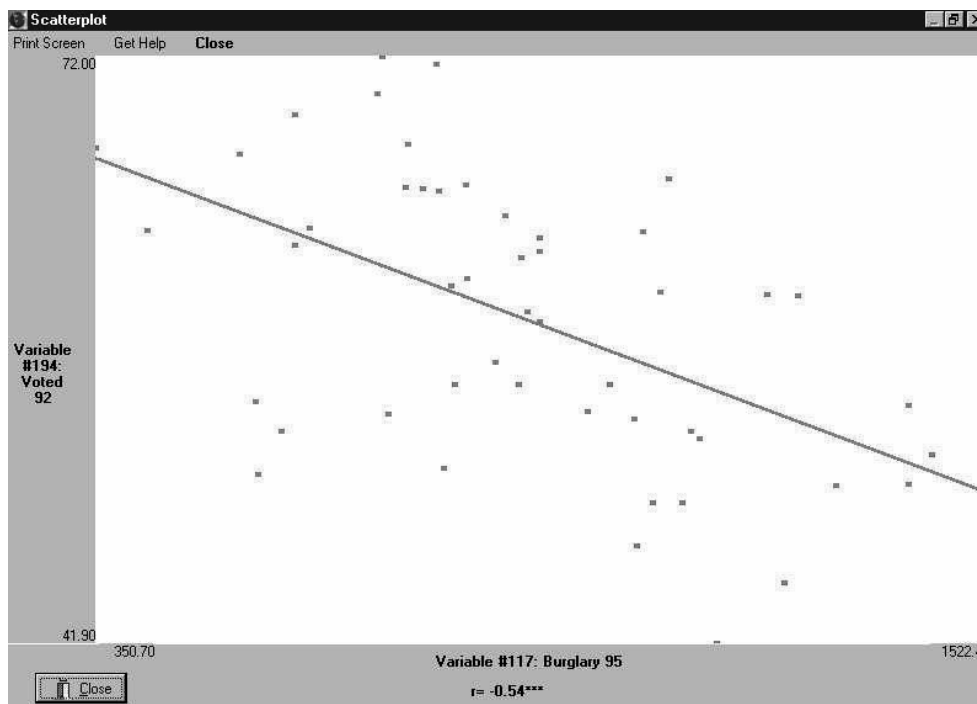
The "**r= 0.61 *****" between burglary and larceny is, in fact, a **strong positive correlation**. Such a high value for "**r**" means that the maps of the variables are essentially the same. Were the "**r**" a bit lower, it would tell us that the maps were pretty much alike, but not completely so. We call this a **moderate positive correlation**. Were the "**r**" lower still, we would know that the maps were somewhat similar, but not overly so. We would call this a **weak positive correlation**.

(I'll be more exact about this later in this chapter.)

NEGATIVE CORRELATIONS AND NO CORRELATION

Now let's look at a negative correlation. **Close** the current scatterplot and **type “117”** (Var 117: BURGLARY 95) as the X-axis variable and "**194**" (Var 194: VOTED 92) as the Y-axis variable. This plots burglary against the percentage of the voting age population that voted in the 1992

presidential election. (Perhaps voters were worried about crime in that election, and you want to find out whether the crime rate dropped as a result.) You should get a graph like this:



We've got a sloped regression line again, but this time it's going down. There seems to be a strong relationship between these variables, but a negative one. The states that are high on voters are low on burglary; those that are high on burglary have a lower percentage of voters. Note the minus sign before the correlation coefficient: " $r = -0.54***$ ". This is a strong **negative correlation**. The scatterplot and the negative " r " show that the maps of these two variables are opposites.

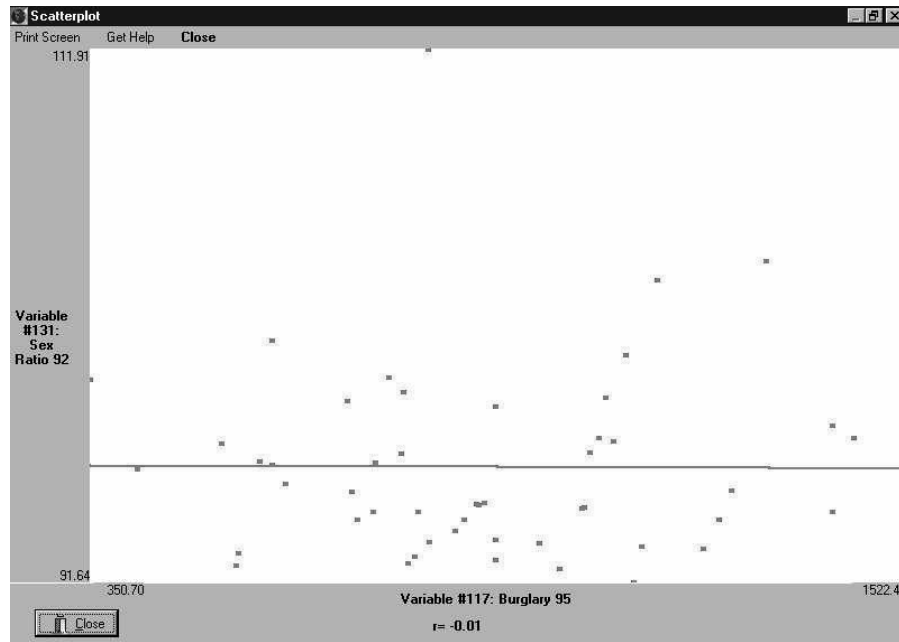
In Chapter 1, we learned that two variables are negatively correlated when they happen in different places. States that are high on the first variable are low on the second. States that are low on the first are high on the second. Places where lots of people voted in 1992 were low on crime three years later, and places where few people voted were high on crime. We don't know if these two things are causally connected, but there is certainly a clear pattern.*

It's a bit harder to display a perfect negative correlation than it is to display a perfect positive one, but *Sociological Insights*[®] can do it. On your own, **close** the previous scatterplot and **type "133"** (Var 133: MOVED 85-90) and **"132"** (Var 132: STABLE 85-90) into the blanks for the X- and Y-axes. The first is the percentage of the population of each state that moved from one house to another between 1985 and 1990. The second is the percentage of each state's population that did not move during those years. These are direct opposites, so the graph on your screen should be the opposite of the graph on page 33. All the dots lie on the line that runs from the top left corner of the chart to the bottom right. Pearson's " r " is "**-1.00*****". This is as low as it can get. Like an " r " of +1.00, an " r " of -1.00 indicates that two variables are perfectly correlated. They are just correlated in different directions.

* No matter how closely two things are correlated, we can not infer that there is a causal connection between them. They may simply happen in the same places (in the case of a positive correlation) or in opposite places (in the case of a negative correlation). Sometimes this is because each of our correlated items is itself caused by a third factor that is correlated with both. We'll see several examples of this in Chapters 4 and 5. Until then, repeat carefully: "Correlation Is Not Cause!"

Again, negative correlations can be **strong**, **moderate**, or **weak**.

If Pearson's "r" runs from +1.00 to -1.00, and both of these show strong correlations, what happens in the middle? What happens when the "r" is close to zero? To find out, **close** this scatterplot and **type "117"** (Var 117: BURGLARY 95) into the blank for the X-axis and **"131"** (Var 131: SEX RATIO 92) into the blank for the Y-axis. You'll get the uncorrelated graph at the top of the next page.



This graph charts the ratio of men to women against the burglary rate in each state. Do areas with a higher proportion of men have more or less burglary than other spots? Apparently not! Rather than forming a pattern, the dots are spread all over the screen. Some states have lots more men than women but are low on burglary; others are high on burglary but have a low male/female ratio; still others are high or low on both. The regression line is pretty much flat: it has no slope to speak of. Pearson's r is -0.01. An "r" of 0.00 means the variables are **uncorrelated***; -0.01 is pretty close to no correlation at all. It looks like the sex ratio and burglary have nothing to do with one another.

There is an important rule here:

positive "r" (up to +1.00)	=	positive correlation (things happen in <u>the same</u> places)
"r" near 0	=	no linear correlation*
negative "r" (down to -1.00)	=	negative correlation (things happen in <u>opposite</u> places)

* The Pearson's "r" measures the direct linear relationship between two variables. There are certain circumstances in which there is a relationship between two variables that is neither direct nor linear. See the section entitled "Linear and Non-Linear Relationships" later in this chapter.

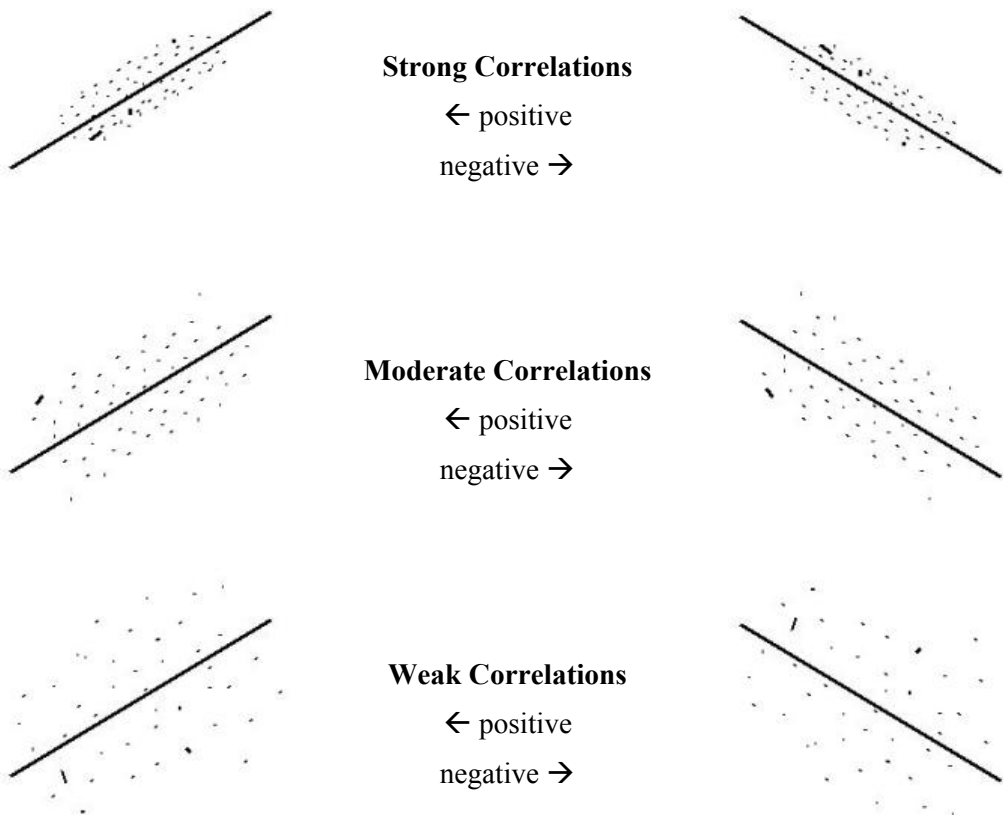
HOW STRONG IS IT?

Okay. An “**r**” of +1.00 is a **strong** positive correlation – indeed, a **perfect** one. The “**r**” of +0.61 between the 1995 burglary and larceny rates is also a strong positive correlation (though not perfect; see page 31). An “**r**” of -1.00 is also a **strong** (and **perfect**) negative correlation, and the “**r**” of -0.54 between the burglary and voting rates is also a strong (but not perfect) negative correlation (page 34). On pages 33 and 35, I noted that both positive and negative correlations can be **strong**, **moderate**, or **weak**. And an “**r**” near zero is no correlation at all.

- What do we mean by **strong**, **moderate**, and **weak** correlations?
- What are the dividing lines between them? Specifically, how do we tell a **weak** correlation from **no** correlation at all?

A full answer to these questions requires more statistical theory than I can provide in an introductory book. But there are partial answers that are good enough for practical work. Let’s take the questions in order.

First, what do we mean by “strong”, “moderate”, and “weak” correlations? Mathematically, the strength of a correlation refers to how closely the dots cluster around the regression line. Take a look at these diagrams:



Strong correlations – either positive or negative – are tightly clustered around the line. Moderate correlations are less tightly clustered. Weak correlations still track the lines, but are more diffuse.

What does this mean in sociological terms? A correlation's strength tells us how easy it is to predict a state's score on one variable if we know its score on another. Imagine that you know where a state falls on the X-axis variable and we want to predict its place on the Y-axis. For which of the above diagrams is our prediction the most likely to be accurate? Clearly, predicting from a strong correlation is more likely to be close than predicting from a weak one. A moderate correlation falls somewhere in-between.

For example, Texas had 1082 burglaries per 100,000 inhabitants in 1995. There is a strong correlation between burglary and larceny, so we ought to be able to predict Texas's larceny rate with some confidence. And we can! Though Sociological Insights[®] doesn't give us the full information, it does tell us that Texas ranks 15th among the states for burglary and 16th for larceny. That's a pretty good prediction. Not all states are that close, of course, but many of them are.

If, on the other hand, we try to predict a state's sex ratio from its burglary rate, we are not apt to be very close. Texas is 13th on one and 15th on the other, but Tennessee is 42nd and 10th, while Montana is nearly the opposite: 10th and 41st. You could "predict" just as accurately by throwing darts at a board. That's because there is no correlation between sex ratios and burglary rates. With no correlation, no prediction is possible.

To summarize: A strong correlation means that we can make such predictions very accurately. A moderate correlation means that our predictions are moderately accurate. A weak correlation means that the predictions are only fairly accurate. When there is no correlation between two variables, it means that we can't make any prediction at all.

Now for our second question: What are the dividing lines between strong, moderate, and weak correlations? This is where it gets tricky, because the answer depends on whom you ask.

Statisticians will tell you that there are no firm dividing lines. Deciding which correlations are "strong", which are "moderate", etc., is a matter of judgment. And they are right! There are no firm lines between them. Each person has to decide how much predictive power makes a correlation "strong", how much makes it "moderate", and – especially – where the line falls between a "weak" correlation and no correlation at all. Mathematically speaking, no dividing lines are possible

Sociologists, on the other hand, don't find this very helpful. We want a rule-of-thumb to guide us – something that doesn't leave everything up to individuals. We especially want to be able to say when a weak "r" indicates is no correlation at all. Most students, I suspect, side with the sociologists. Among other things, it's pretty nerve-wracking to have to guess where your instructor draws the line between the various types of correlations, etc., so you won't flunk your exams. Fortunately, a rule-of-thumb is available. Even better, your computer gives it to you.

Take a look at the "r" value at the bottom of any scatterplot screen. In most cases, that "r" is followed by one, two, or three asterisks (*, **, ***). The strong positive correlation between Var. 117: BURGLARY 95 and Var. 118: LARCENY 95 is shown as 0.61***; the strong negative correlation between Var. 117 and Var. 194: VOTED 92 is shown as -0.54***. Sometimes there aren't any asterisks: Var 117 and Var 131: SEX RATIO 92 are uncorrelated and their "r" is shown as a plain -0.01. These asterisks serve as the rule-of-thumb that sociologists want. As a rule-of-thumb:

***	=	very important ("strong") correlation
**	=	moderately important ("moderate") correlation
*	=	weak, but still important ("weak") correlation
(none)	=	<u>no</u> correlation at all

Notice the language in this last box: it refers to “important” correlations, “moderately important” correlations, and “weak but still important” correlations. It then identifies each of these with the terms “strong”, “moderate”, and “weak”. This is because the rule-of-thumb is based on a mathematical deception. The asterisks that sociologists use actually have nothing to do with the strength of a correlation, but sociologists have discovered that we can treat them as if they do. Let me explain.

On pages 1 and 2 of this book, we learned the difference between aggregate data and survey data. We learned that the first of these is based on data covering an entire population (in this case, all of the 50 states), while the second is based on a sample from a population. Surveys take data from a limited number of people and generalize their results to the population as a whole. The General Social Survey – which is the basis for Sociological Insights[®] calculations – polls about 3000 people every other year, from which it draws conclusions about the entire (or almost entire) American adult population.

We’ll learn much more about this in the second half of this book. For now, we simply need to know that there are statistical tests that tell us how likely it is that the answers we get from our sample reflect the views of the population as a whole. These are called “tests of statistical significance”. The asterisks that accompany our correlations actually test the **statistical significance** of those correlations, not their strength.* That is, they treat our aggregate data as if it had come from a sample, then measure how likely it would be that a correlation in that sample would reflect a correlation in a (hypothesized) population. Sociologists read the results as if they were a measure of strength – with clear dividing lines.

- A correlation of 0.60 in a sample of 50 is almost certain to reflect a real correlation in a population of whatever size. In fact, the odds are 999 to 1 that it does so. Sociologists reason that only strong correlations are likely to do this, we call a three-asterisk correlation “strong”.
- A correlation of 0.45 in the same size sample is highly likely to reflect a real correlation in the underlying population, with odds of 99 to 1. Sociologists reason that moderate correlations are likely to do so, so we call a two-asterisk correlation “moderate”.
- A correlation of 0.30 in the same size sample very likely to reflect a real correlation in the underlying population, this time with odds of 19 to 1. Sociologists accept such correlations as “weak”.
- Any correlation that produces odds of less than 19 to 1 is treated as if it were no correlation at all. (As we’ll see later on, sociologists insist on having good odds that our conclusions are right. If a correlation has less than a 1 in 20 chance of being right, then we say that there is no correlation at all.)

This is all quite reasonable, even though it drives mathematical statisticians nuts. It has now become standard sociological practice to treat a correlation’s statistical significance as a stand-in for that correlation’s strength. This provides clear dividing lines between the categories. It also makes our work easier, because nearly every statistical program produces these asterisks automatically. Sociology books and journal articles now regularly use the number of asterisks beside a correlation to differentiate between “strong”, “moderate”, and “weak” correlations. We especially use them to separate these levels of correlation from no correlation at all.

The rest of this book will follow this sociological convention. We will use asterisks as a rule-of-thumb to measure correlations’ strength. We will call three-asterisk correlations “strong”,

* Thanks to Gary Schulman and Rae Newton for pointing this out to me.

two-asterisk correlations “moderate”, and one-asterisk correlations “weak”. We will do so without apology, but recognizing that doing so is not technically correct.

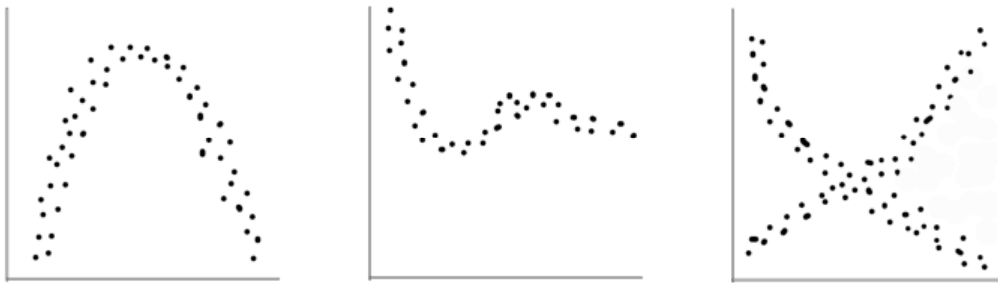
Particularly, we will interpret no asterisks as indicating no correlation. This gives us a clear way to determine which relationships between our variables are important. We need that, if we are to make any progress.

Though technically problematic, this rule-of-thumb is the best choice available.*

LINEAL VS NON-LINEAR RELATIONSHIPS

As you’ve undoubtedly noticed, all the lines in the preceding graphs are straight. That’s because the Pearson’s “r” only measures how closely the dots cluster around the particular straight line that comes the closest to connecting the dots together. If the dots are spread all over the screen, the Pearson’s “r” assumes that there is no relationship between them. This is true! – but only if we limit ourselves to straight lines.

Some scatterplots, however, are different. Take a look at the following:



Each of these graphs shows a clear relationship between the X and Y variables, but none of these relationships is a straight line. In the first and third cases, Pearson’s “r” will be close to 0, indicating no correlation, even though there is one. The middle graph will likely show a weak negative correlation, but even that misses the close relationship between the variables.

Scatterplots like these are beyond the scope of this book – or, indeed, of all but the most advanced statistics texts. The relationships you will encounter here are much simpler, as are the relationships most sociologists encounter in their research. It is always worth checking your scatterplots to see whether some strange relationship lies hidden in your data. For the most part, however, the Pearson’s “r” will give you all the information you need.

THE CORRELATION MATRIX

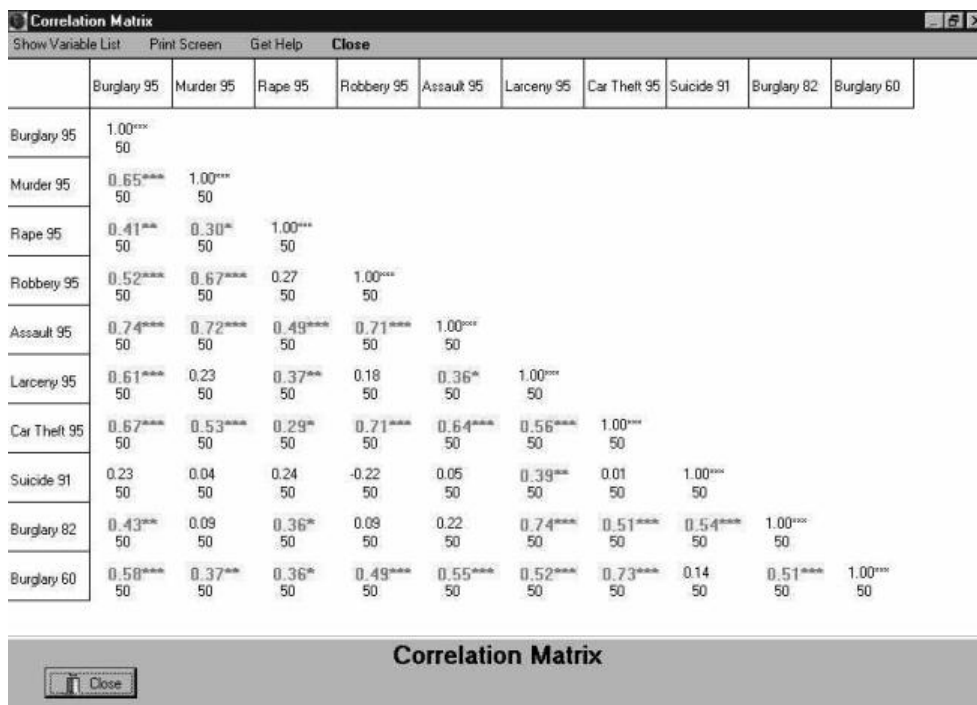
As you’ve probably discovered, making a scatterplot for every pair of variables you want to look at takes time – even on the computer. [Sociological Insights®](#) allows you to take a shortcut: you can figure correlation coefficients on a group of variables all at once. Here’s how:

* In fact, this choice is less problematic with a 50-state data set than with something larger. The logic of sampling gives the same level of correlation more or fewer asterisks, depending on the size of the sample. A “weak” correlation of 0.25 in a 50-unit sample becomes a “strong” correlation in a 3000-unit sample. Sticking to a 50-state sample not only makes significance and strength more congruent; it also allows us to compare the correlations between various pairs of variables without worry. A 0.30 correlation between two of our variables is certainly weaker than a 0.60 correlation.

Close the scatterplot, then click the “Cancel” button to get back to the main program screen. Click the **S**Tates menu and choose “Correlate”. Then type the following as variables 1-10, pressing “<TAB>” between each one:

117, 112, 114, 115, 116, 118, 119, 239, 227, 228

Click the “OK” button to get this screen:



	Burglary 95	Murder 95	Rape 95	Robbery 95	Assault 95	Larceny 95	Car Theft 95	Suicide 91	Burglary 82	Burglary 60
Burglary 95	1.00*** 50									
Murder 95	0.65*** 50	1.00*** 50								
Rape 95	0.41** 50	0.30* 50	1.00*** 50							
Robbery 95	0.52*** 50	0.67*** 50	0.27 50	1.00*** 50						
Assault 95	0.74*** 50	0.72*** 50	0.49*** 50	0.71*** 50	1.00*** 50					
Larceny 95	0.61*** 50	0.23 50	0.37** 50	0.18 50	0.36* 50	1.00*** 50				
Car Theft 95	0.67*** 50	0.53*** 50	0.29* 50	0.71*** 50	0.64*** 50	0.56*** 50	1.00*** 50			
Suicide 91	0.23 50	0.04 50	0.24 50	-0.22 50	0.05 50	0.39*** 50	0.01 50	1.00*** 50		
Burglary 82	0.43*** 50	0.09 50	0.36* 50	0.09 50	0.22 50	0.74*** 50	0.51*** 50	0.54*** 50	1.00*** 50	
Burglary 60	0.58*** 50	0.37** 50	0.36* 50	0.49*** 50	0.55*** 50	0.52*** 50	0.73*** 50	0.14 50	0.51*** 50	1.00*** 50

This screen shows us a **correlation matrix**. It lists each variable both across the top and down the left side. If you pick a top variable and read down the column below it, you see the correlation coefficient for each pair. (You also see the number of states on which we have information for both values – 50 for each of these pairs.) Reading down the leftmost column, for example, we’ve got coefficients between each variable and the burglary rate in 1995. The first few are the murder, rape, robbery, assault, larceny, and car theft rates for the same year. Then comes a coefficient between the 1995 burglary rate and the suicide rate a few years earlier (that’s the closest data we have). Last come coefficients between the 1995 burglary rate and burglary rates 13 and 25 years earlier.

What do these coefficients mean? In the first column, all but one of the correlations are **positive**. And all but three of these positive correlations are **strongly** positive. (We know this because they have three asterisks by them indicating large values for Pearson’s “r”.) In 1995, burglary, murder, robbery, assault, larceny, and car theft all happened in the same places. And burglary in 1960 also happened in those places. One crime may not cause another, but something is going on in those places that pushes all of these crimes up.

Burglary in 1995 is only **moderately** correlated with rape in that year, and also only moderately correlated with burglary in 1982. This means that the match is still present, but not as strong. (It is still important, however.) The two asterisks by these r’s tell us this.

The lack of asterisks by suicide, and its low r-value (0.23), tell us that suicide and burglary don’t happen in the same places in the 1990s, and they don’t happen in opposite places either. **There is no direct relationship between these variables.** Read that sentence again. It’s not that we can’t

tell what the relationship is between these two variables, because we can. **They have no direct linear relationship**. That's what being uncorrelated means.

Note that none of these correlations is negative. **Negative** correlations indicate a relationship between two variables, just as **positive** ones do. But the **negative "r"** means that things happen in opposite places, while the **positive "r"** means that things happen in the same places. It's very important to keep this straight.

Sociological Insights[®] helps us out here by highlighting the significant correlations in red. If an r-value is black, the two variables it represents are uncorrelated.^{*} If an r-value is red, the two variables are correlated. You have to look for the minus sign to see whether the correlation is positive or negative. But you can tell which pairs are correlated and which are uncorrelated at a glance.

positive correlations mean that things happen in the same places
negative correlations mean that things happen in opposite places
no correlation means that they don't happen in either the same or opposite places – there is no direct linear relationship between the variables

A TIMELY WARNING

We'll deal with the topic at greater length in the next chapter, but it is worth taking a moment here for a timely warning.

- We have seen that states with high burglary rates also have high larceny rates. This does not mean that burglars commit financial fraud.
- We have seen that states in which a high percentage of the population voted in 1992 have lower burglary rates. This does not mean that voters don't steal!

In the first case, it is possible that no one – not a single person – has committed both burglary and larceny. We just have high rates of both in some places and low rates of both in others. In the second case, all the thieves could vote and all the non-voters could all be honest. It's just that places with lots of voters don't have much burglary and places with lots of burglary don't have many voters.

High or low rates of something in a given area don't tell us which individuals are doing that thing.

Claiming that the rate at which things happen in a given area tells us something about the actions of specific individuals has a name in sociology. It is the **ecological fallacy**. Such claims don't work. Data about areas doesn't tell us anything about individuals. It is important to avoid this fallacy as one works with aggregate data.

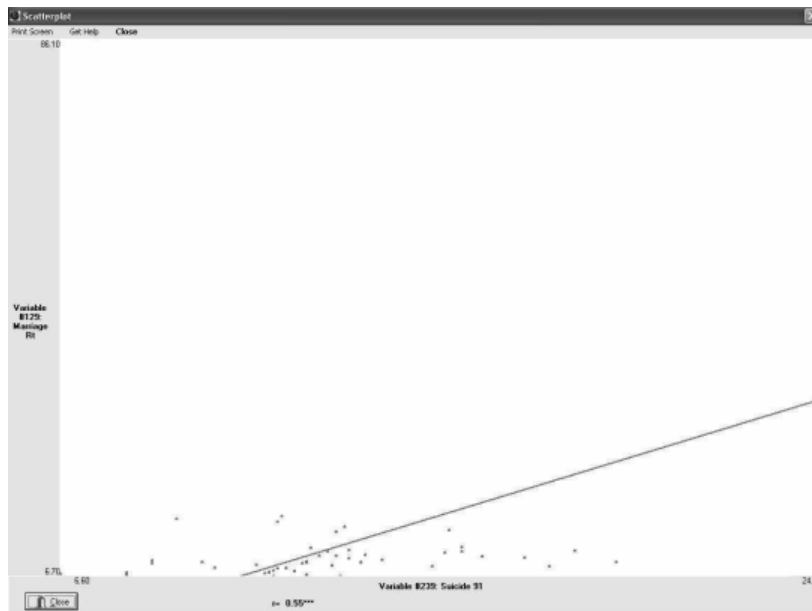
^{*} The correlation of a variable with itself is also black. That's because correlating a variable with itself always gives an "r" of 1.00. This is not news. Sociological Insights[®] reserves red for things that are worth viewing.

OUTLIERS

There's one more thing to watch out for when working with correlations. In the last chapter, we saw that the median is sometimes a better measure of a variable's central tendency than is the mean. This is true when one or two states are so much different from the rest that they pull the mean way up or way down. Such **outliers** (see page 19) distort the overall pattern.

Correlations can also have outliers, which are just as misleading here as they are elsewhere. Their extreme values can shift the Pearson's "r" so much that they can create the illusion of a correlation that is not really real. The test is to take a look at the scatterplot. If we find one or two states that are widely separated from the others, we need to give our raw correlation figures a closer look.

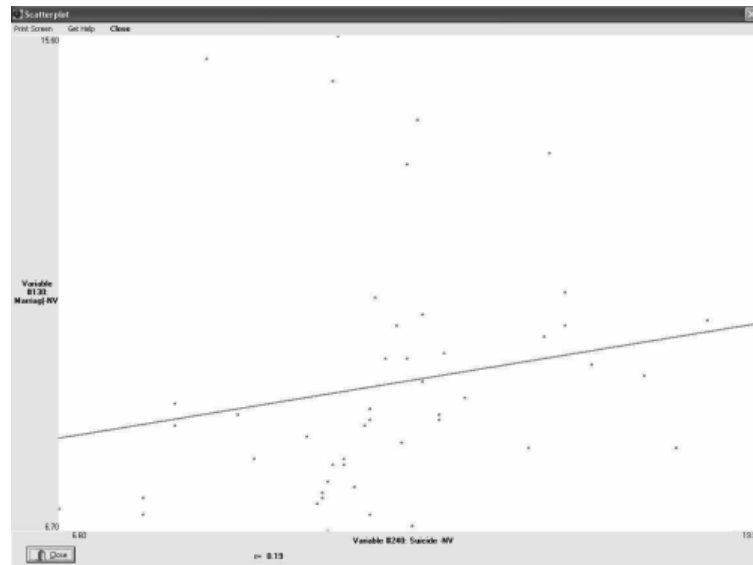
Here's an example. Choose "**Scatterplot**" from the **STates** menu, then use "**239**" (the suicide rate in 1991) as the X-axis variable and use "**129**" (the number of marriages per 1000 population in 1992) as the Y-axis variable. Click "**OK**" to show the scatterplot (found at the top of the next page).



As you can see, most of the states are clustered toward the bottom of the graph. Way up in the right-hand corner, however, is a single dot – Nevada. Nevada is a classic outlier on both of these variables. Its suicide rate of 24.8 per 100,000 population is 20% higher than Montana's, the next most suicide-prone state. Its marriage rate of 86.6 per 1000 population is 5½ times the rate in the next most marriage-happy state, Arkansas. Nevada has long been known as a place for quickie marriages and divorces. It appears to be a popular place for suicide as well.

Now take a look at the Pearson's "r". It is a strong 0.55 (three asterisks). This would seem to indicate that suicide and marriage are related to each other. But the rest of the states, besides Nevada, don't seem to show this pattern. We need to eliminate Nevada, if we are to see whether the correlation between suicide and marriage is a general pattern or the result of a single wild case..

Close this scatterplot, and this time use "**240**" for the X-axis variable and "**130**" for the Y-axis variable. These are the same variables, with Nevada taken out.



As you can see, the dots are all over the map. The regression line still rises from left to right, but the Pearson's "r" has dropped to 0.19. This is not high enough to indicate a correlation between the suicide rate and the marriage rate. The fact is, Nevada alone created our previous correlation.

One should always check one's correlations against their scatterplots, looking for such cases. And one should always eliminate outliers this extreme.*

THINGS TO REMEMBER

1. **r is positive** ⇔ positive correlation
 - ⇔ as the rate of X increases, the rate of Y increases, or vice versa
 - ⇔ the regression line slopes upwards from left to right
 - r close to 0** ⇔ no direct linear correlation
 - ⇔ an increase or decrease in either variable is not related to an increase or decrease in the other
 - ⇔ the regression line is flat from left to right
 - r is negative** ⇔ negative correlation
 - ⇔ as the rate of X increases, the rate of Y decreases, or vice versa
 - ⇔ the regression line slopes downwards from left to right
2. Remember our rule-of-thumb:

r ***	⇔ very important (“ strong ”) correlation
r **	⇔ important (“ moderate ”) correlation
r *	⇔ less important (“ weak ”) correlation
r (no stars)	⇔ no correlation

* Sociological Insights[®] can't eliminate outliers in calculating correlations. Research-oriented statistics programs like SPSS and Microcase generally allow one to do so, usually by creating new variables that leave out extreme cases. Even then, however, you have to check the scatterplots to see which cases to remove.

(This rule is *technically* incorrect, because the computer's asterisks actually measure statistical significance. Sociologists routinely use them as a stand-in for strength, however.)

3. The fact that two variables are **uncorrelated** does NOT mean that we can't tell what the relationship is between them. We can. Uncorrelated variables have no direct linear relationship between them!
4. As we saw in Chapter 1: **Correlation does not (necessarily) mean cause**. Two variables may be correlated – either positively or negatively – because they are both correlated to a third. Their apparent relationship may thus be a product of their independent relationship with that third variable. (We will pursue this idea in Chapter 4.)
5. The **ecological fallacy** involves drawing conclusions about individuals or classes of people solely based on events happening in their areas.
6. Watch out for **outliers** – states that are so different from the rest that they can sometimes create a correlation all by themselves. It's always best to check each correlation against its scatterplot. This makes sure that we're seeing overall patterns, not exceptions.
7. The Pearson's "r" measures the **direct linear relationship** between two variables. It does not pick up non-linear relationships. The advanced statistical tests that would pick up these non-linear relationships are beyond the scope of this book.

FOUR: MULTIPLE REGRESSION (I)

In the last chapter, we studied correlations. These measure the degree to which two things "co-relate" or vary together. The burglary and larceny rates are positively correlated: both are high in some states, and both are low in others. Burglary and voting rates are negatively correlated: where one is high, the other is low. Burglary is not correlated with a state's sex ratio; there is no direct linear relationship between these two variables.

Of courses, these are generalizations, which may not hold true for all eras. Though burglary and larceny are consistently found in the same places (at least in 1995, 1982, 1960, 1940, and 1923), burglary and murder are not. They are sometimes correlated (1995, 1940, 1923) and sometimes uncorrelated (1982, 1960). In either case their relationship is a social fact that needs to be explained. And sociologists are supposed to explain facts like this.

As sociologists, we want to go beyond the simple fact that things are correlated; we want to know why they are. **What is the underlying reason that things vary together?** Logically, there are three possibilities.

First (and easiest), **one thing may cause another**. For example, the high rate of violent crime in a given area may lead people to hire more police officers: this would explain why the correlation between Variable 25: VIO CRIME 95 and Variable 60: COPS/10K 92 is +0.52. The high crime rate causes larger police forces.*

Likewise, a high percentage of church goers in some states may contribute to a lower crime rate, perhaps because of effective sermons, but more likely because people don't want to lose the good esteem of their fellow religionists. (A negative correlation can indicate a causal relationship just as much as can a positive one.)

Note that social facts may have more than one cause. Larger police forces may be not just a consequence of higher crime rates. They may also depend on a community's wealth and its willingness to spend that wealth on public services. Take a look at the following correlation matrix:

	Cops/10K 92	Med \$ 90-92	Teach \$ 95
Cops/10K 92	1.00*** 50		
Med \$ 90-92	0.31* 50	1.00*** 50	
Teach \$ 95	0.35* 50	0.75*** 50	1.00*** 50

The number of police per ten thousand population (Var 149: COPS/10K 92) correlates with a state's median income (Var 138: MED \$ 90-92) at +0.31. It also correlates with teacher salaries (Var 198: TEACH \$ 95) at +0.35 – a measure of how much public money a state's citizens were

*Technically, we should use an earlier figure for violent crime: the violent crime rate in 1995 logically cannot influence the number of police employed in 1992. But we don't have figures for the later year, so we have to use what we've got. Hiring more police probably does not increase the crime rate, so we don't have to worry about getting the causality backwards. And since neither rate varies much over such a short time, we lose little by using the figures that we have available.

willing to spend at that time. These factors – along with crime – may together increase the number of police. Again, the operative word is “may”.

But correlation does not necessarily imply cause! A second possibility is that the **two variables may measure essentially the same thing**. The burglary rate and the larceny rate in 1982, for example, correlate at +0.82. This is a huge correlation for sociology. However, it makes no sense to speak of burglary "causing" larceny, nor vice versa. Think of both rates as measuring an underlying variable: “crimes of property”. Where “crimes of property” are high, the burglary and larceny rates will both be large. Where there is a lot of larceny, there will be a lot of burglary, because the crime rate in that area is high.

Likewise the high correlations between Var 132: STABLE 85-90 and Var 133: MOVED 85-90 arise because both measure the same thing -- the degree to which a state's population stays in one place. The former is the positive measure of this (how many people don't move); the latter is the negative (how many people do move). STABLE and MOVED are perfectly negatively correlated with each other, but they are still measuring the same thing. (Run a scatterplot for these two variables to see their exact relationship.)

There is also a third possibility. **Two variables may both be caused by a third**. They may appear to be correlated simply because they both co-occur with a third variable that is really causing them both. This happens a lot, as we'll see in this chapter and the next.

Three causal possibilities:

1. A & B are correlated because A causes B.
2. A & B are correlated because they measure the same thing.
3. A & B are correlated because they are both caused by C.

In the first case, A is the **independent variable** and B is the **dependent variable**. In the second case, neither variable is either independent or dependent. In the third case, C is the real independent variable and both A and B are dependent on it.

All by itself, then, correlation tells us nothing about cause. Just because two things happen in the same places does not let us conclude that one causes another.

To explore causality, sociologists use a technique called **multiple regression**. Multiple regression can tell us two things.

- First, it can separate those cases in which a real relationship between two variables produces their correlation (#1 in the box above) from those cases where the only connection between two variables is that both are caused by a third (#3 in that box).
- Second, whenever two or more variables do cause a third – independently of one another – regression can tell us the relative contribution that each one makes.

Let's look at an example. **Start Sociological Insights[®], click the STates menu, choose “Correlate”, and type the following variables in the blanks:**

240, 133, 155, 160, 166, 171, 186, 201, 131 and 196

Click on “OK” to get the correlation matrix that appears at the top of the next page. This matrix correlates the suicide rate in 1991 with a series of other variables. (Note that this particular suicide rate does not include Nevada. As we saw at the end of the last chapter, Nevada has a much higher rate of suicide than does any other state. It is an **outlier**, whose very unusualness is likely to

prevent us from seeing the overall pattern. It's best to leave it out.)

	Suicide -NV	Moved 85-90	BreastCanc90	No Ins 90-92	Car Death 92	School \$ 92	% Kids 92	Pickups 89	Sex Ratio 92	Commute 90
Suicide -NV	1.00*** 49									
Moved 85-90	0.49*** 49	1.00*** 50								
BreastCanc90	-0.49*** 49	-0.39*** 50	1.00*** .50							
No Ins 90-92	0.34* 49	0.35* 50	-0.26 .50	1.00*** 50						
Car Death 92	0.66*** 49	0.13 50	-0.42** .50	0.59*** 50	1.00*** 50					
School \$ 92	-0.47*** 49	-0.13 50	0.48*** .50	-0.46*** 50	-0.60*** 50	1.00*** 50				
% Kids 92	0.48*** 49	0.25 50	-0.52*** .50	0.25 50	0.38** 50	-0.41** 50	1.00*** 50			
Pickups 89	0.72*** 49	0.25 50	-0.43** .50	0.27 50	0.69*** 50	-0.56*** 50	0.60*** 50	1.00*** 50		
Sex Ratio 92	0.31* 49	0.72*** 50	-0.52*** .50	0.02 50	0.04 50	0.08 50	0.50*** 50	0.34* 50	1.00*** 50	
Commute 90	-0.47*** 49	-0.07 50	0.28 .50	0.18 50	-0.39** 50	0.28 50	-0.40** 50	-0.70*** 50	-0.32* 50	1.00*** 50

What do we find? Looking down the leftmost column, we see that all of the correlations with the suicide rate are significant.* Some are strongly so, others only moderately. States where lots of people drive pickup trucks and there are a lot of traffic deaths have high rates of suicide, as do states where there is lots of residential mobility, high numbers of people without health insurance, more men than women, and large numbers of children. Is living around a lot of excess males hazardous to your health? Does driving where there are lots of pickup trucks lead one to commit suicide? It's easy to see how having lots of pickup trucks could lead to traffic deaths – those two variables' correlation is +0.69 – but it's hard to see how both are connected to suicide. But the +0.72 and +0.66 correlations are very strong.

On the other hand, suicides are lower in states with higher rates of breast cancer and higher spending on schools, just as they are lower in states with higher urban populations, with longer average commutes, with more Catholics and Jews (Var 190: % CATH 90; Var 192: % JEWISH 94), and with high teacher salaries (Var 198: TEACH \$ 95). Does this mean that spending money on teachers and schools lowers the suicide rate? Could we eliminate suicide by moving everyone to cities, changing their religion, taking away their health insurance and giving them breast cancer?

Correlation does not necessarily indicate cause!

Finding causal relationships requires more analysis.

Some of these correlations show real relationships between the variables. Others are what

* Clever readers will have noticed that all of these variables are from the early 1990s. This is because we have a wider choice of variables from that era than from others. It doesn't make much sense to argue that having no health insurance in 1990 will raise the suicide rate ten years later – so we use variables from the same era.

sociologists call **spurious**. Their high **Pearson's coefficient (r)** is real enough – but it doesn't indicate a real connection between the variables. Both variables are caused by a third. Sociologists use **multiple regression** to separate the real correlations from the spurious ones. We'll now see how it's done.

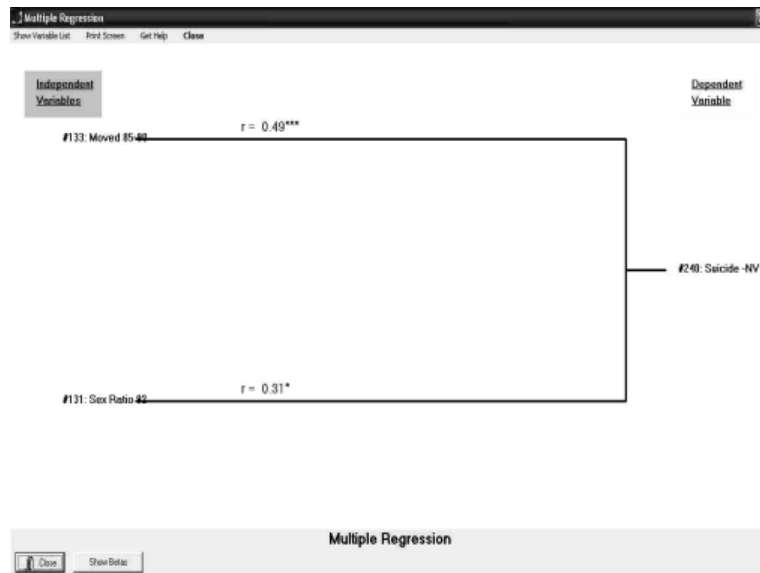
EXPOSING SPURIOUS CORRELATIONS

Before we start, I need to emphasize that regression analysis requires a bit of thinking. One can't sort true correlations from spurious ones mechanically. We have to begin somewhere, and the best place to begin is with patterns that sociologists have previously found to be important. We can't, for example, use the relationship between pickup trucks and suicide to sort the truth or falsity of the rest of the correlations we have just seen. We need to pick a variable that more plausibly influences both the suicide rate and other matters.

Spurious means that despite the high Pearson's coefficient (r), the two variables have no real relationship with each other. They are instead both caused by a third.

In this case, we have one to hand. Variable 133: MOVED 85-90 reports the percentage of each states' population that moved from one dwelling to another between 1985 and 1990. Sociologists have previously discovered that high rates of residential mobility are tied to various social disorders. People seem to do different things in places where people move around a lot and other things in places where people are relatively stable. MOVED 85-90 does not have the highest correlation with SUICIDE -NV in our set, but its "r=0.49***" is very respectable. We'll use this to begin to separate the wheat from the chaff.

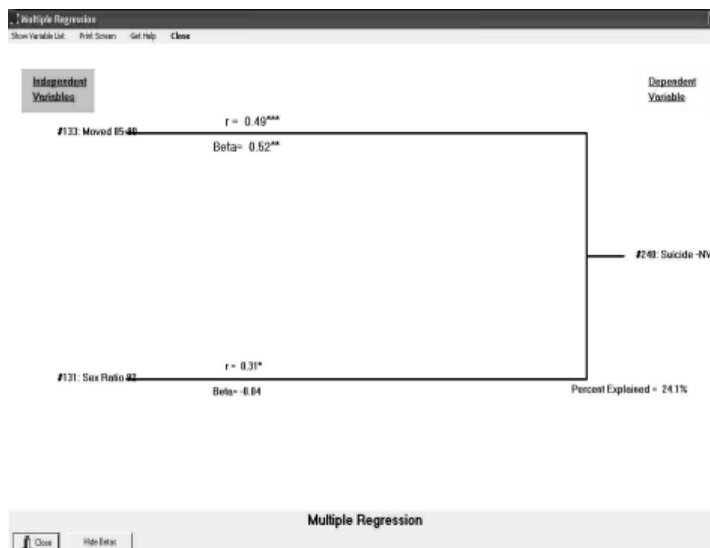
Hit <ESC> to go back to the *Sociological Insights*[®] opening screen. From the **S**States menu, **click on "Regress"**. **Type "240"** (Var 240: SUICIDE -NV) as the Dependent Variable, and **type "133"** (Var 133: MOVED 85-90) and **"131"** (Var 131: SEX RATIO 92) as two Independent Variables. **Click "OK"** to get the following:



What on earth is this? It is a **path diagram** showing the proposed connections between **two independent variables** on the left – the various states' residential mobility and their ratio of men to women – and **one dependent variable** on the right – the suicide rate (leaving out Nevada). Just above each of the horizontal lines in the path we find the correlation coefficient ("r=") between each independent variable and the dependent variable. (Remember that independent variables are the ones that we think might be influencing the dependent ones.)

As you can see, a state's sex ratio correlates with that state's level of suicide at a +0.31 rate. The percentage of the population that moved in the 5 years before the 1990 census correlates with suicide at a rate of +0.49. Both correlations are positive. States with lots more men than women, and states where lots of people move each year, both have higher levels of suicide. So far this is nothing new; it just puts into graphic form what we already know.

Now **click the “Show Betas” button** at the bottom of the screen. You'll get the following:



This screen is like the last one, except for four things. Taking the least important first, the “Show Betas” label has changed to a “Hide Betas” label. That way, you don’t have to start all over again if you want to revisit the previous screen.

In addition: there is now a new number below each horizontal line, saying “Beta = ____”. There is a “Percent Explained = ____” in the lower right hand corner. And the Pearson’s r for SEX RATIO has changed color from red to black. We’ll explain each of these three in turn.

When two factors each cause a third, there is almost always some interaction between them. Let's assume that there is something about the places where there are lots more men than women that makes people prone to suicide, and also something about places where lots of people move often that also disposes people to suicide.* There will be some people in these areas who will be led to suicide by being around too many men, and some will be led to it by being around people who have moved. And there will be some who will be led to it by being around both. In order to find out how influential each factor is, we have to find a way to measure them as if they were independent of one another.

Multiple regression does exactly this. It asks, in effect, “What would be the relationship between the sex ratio and the suicide rate, if we eliminate the influence of people moving?” We call this **controlling** for residential mobility. Essentially, we get the computer to pretend that the same percentage of every state’s population moved in the last five years. What kind of connection

* Remember that we don’t want to commit the **ecological fallacy** (see pages 33 and 54). A high correlation between the suicide rate and a high ratio of men to women does not mean that single men are committing suicide. All the suicides might be by women. Neither does the connection between residential mobility and suicide mean that people who move are more prone to killing themselves. It might be the stay-at-homes doing it. The most we can say is that something in the social climate of places where there are a lot more men than women and places where there is lots of residential mobility leads to increased rates of suicide.

is left between the sex ratio and suicide, once we have taken residential instability out of the picture? The resulting figure is called the **Beta**: it appears just below the "r=" on the line connecting SEX RATIO with SUICIDE. It is -0.04.

The Beta on the line connecting MOVED with SUICIDE arises from a similar question: "What would be the relationship between the percentage of the population that has moved and the suicide rate, if we eliminate the influence of an imbalance between men and women?" This time, we're **controlling** for the state-to-state variation in sex ratio. Again, the computer pretends that all states have equal sex ratios, and isolates the independent impact that residential instability has on the suicide rate. The resulting Beta is +0.52**, almost the same as the "r=" that appears right above it.

What do these **Betas** mean? Technically, they are the **net effect** that each independent variable has on the dependent variable, expressed in standard units.* We read them just like Pearson's "**r**". A starred positive Beta on the independent variable means that there is a significant positive relationship between that variable and the dependent variable, after the effects of the other independent variable(s) have been removed. A starred negative Beta on the independent variable means that there is a significant negative relationship between that variable and the dependent variable, after the effects of the other independent variable(s) have been removed. A large Beta means a strong relationship (positive or negative). A moderate Beta means a moderate relationship.

Sociological Insights[®] uses the same system of asterisks for **Betas** as it does for **r**'s: three asterisks (***) indicates a strong relationship, two (**) indicates a moderate one, and one (*) indicates a weak one.** Comparing the size of Betas allows us to compare the relative influence that each independent variable has on the dependent variable – either positive or negative. The +.52 is much larger than the -.04, telling us that MOVED has much more influence on SUICIDE than SEX RATIO does.

(Remember to ignore the plus and minus signs: Betas of +0.52 and -0.52 are equally strong, as are Beta's of -0.04 and +0.04. Like Pearson's "r", Betas get stronger, the farther they are from zero in either direction.)

In fact, if an independent variable's Beta is small enough – if it has no asterisks after it – there is no real relationship between that variable and the dependent variable. It means that the original correlation was not the result of a real connection between the variables but was a by-product of the influence of the other dependent variable. The original correlation was **spurious**.

Regression analysis tells us the separate influence of each independent variable on the dependent variable, after **controlling** for the influence of all the other independent variables.

(This is also sometimes called "**holding** the other independent variables **constant**" or using them as "**control variables**".)

* The standard units are **z-scores**, a method of scoring that allows us to compare the relative influence of variables that have different units of measurement. One figures z-scores by dividing the variable's actual score by its standard deviation. This lets one compare apples and oranges, at least to the degree that each varies from its mean.

Statisticians warn that standardized Betas can vary greatly from sample to sample – much more than do the unstandardized regression coefficients (usually symbolized by "**b**"). We do not have to worry about this with our States data, because it comes from an entire population, not a sample. We would expect, however, the population of the 3000+ U.S. counties to produce somewhat different results.

** The same caveats apply here that apply to the use of asterisks to measure the strength of correlations. See page 38.

And that is just what we find in this case. The Beta between SEX RATIO and SUICIDE is -0.04: there is no relationship here. The Beta between MOVED and SUICIDE is +0.52, which is very large. The original Pearson's "r" that connected the sex ratio with suicide was spurious: a result of the fact that people in stable areas don't (tend to) commit suicide, and those areas also don't have a lot of excess males. A sex ratio skewed toward men and the suicide rate are both related to the same thing: residential instability.

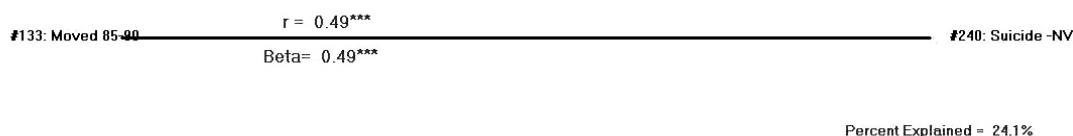
red = real
black = spurious

This is where the color shift comes in. If a Beta is high enough to indicate a real relationship between the independent and dependent variables, *Sociological Insights*[®] displays it in **red**. If, on the other hand, the Beta is so low that it indicates no real connection between the variables, *Sociological Insights*[®] displays it in **black**. When you push the "Show Betas" button, you can see at a glance which variables are **spurious** and which are **not spurious** (real). The Betas for spurious variables will be in black; those that show a real connection between the variables will be in red.

The regression analysis gives one last bit of information. Look at the number in the lower right hand corner of your screen. It says "**Percent Explained = 24.1%**". This means that the two independent variables – taken together – explain 24.1% of the variation of the dependent variable. And because SEX RATIO does not contribute anything to this figure – its Beta is too low – that variation comes from the variation in MOVED.

What do we conclude? That we can explain nearly 24% of the state-to-state-variation in the suicide rate by the variation in the rate of residential instability. Where people move more often, there are more suicides. We can understand this, because suicides are often people who feel no connection to others. Even if it is not the people who move who commit suicide, social connections are weaker where people move more often, so it make sense that the suicide rate would be higher there.

Close this regression and start a new one, again with "**240**" as the dependent variable and with "**133**" as the first independent variable. This time, however, don't enter any more independent variables. Instead, **click "OK"**, then **click "Show Betas"** to get the following path diagram:



This is a multiple regression with only one independent variable: MOVED. As such, it's not too useful. But it does tell us how much of the variation in SUICIDE this particular independent variable explains. And look at that: **Percent Explained = 24.1%**. It's almost exactly the same as on the previous screen, when we had two independent variables. Using SEX RATIO as a second independent variable added nothing! This confirms the previous analysis.

In general, we don't run multiple regressions with just one independent variable; there is a reason for the word "multiple", after all. Let's now take up another example.

TWO SHORT EXAMPLES

Example 1: Do Large Concentrations of Catholics Increase the Abortion Rate?

Everyone – or almost everyone – knows that the Roman Catholic Church opposes abortion. As part of its "consistent life ethic," the Church regards abortion as the taking of a human life. Given that Catholicism has more members than any other church in America, it stands to reason that there should be fewer abortions in states where there are more Catholics. That is, we should expect that

Multiple Regression (I)

the percentage of the population that is Catholic (Var 190: % CATHOLIC) and the abortion rate per 1000 live births (Var 185: ABRT/BIR) should be negatively correlated. States like Rhode Island, Massachusetts, and Connecticut – where about half the population is Catholic – should have fewer abortions per 1000 births than states like Alabama, Tennessee, and Arkansas, where less than 5% of the people are Catholic.

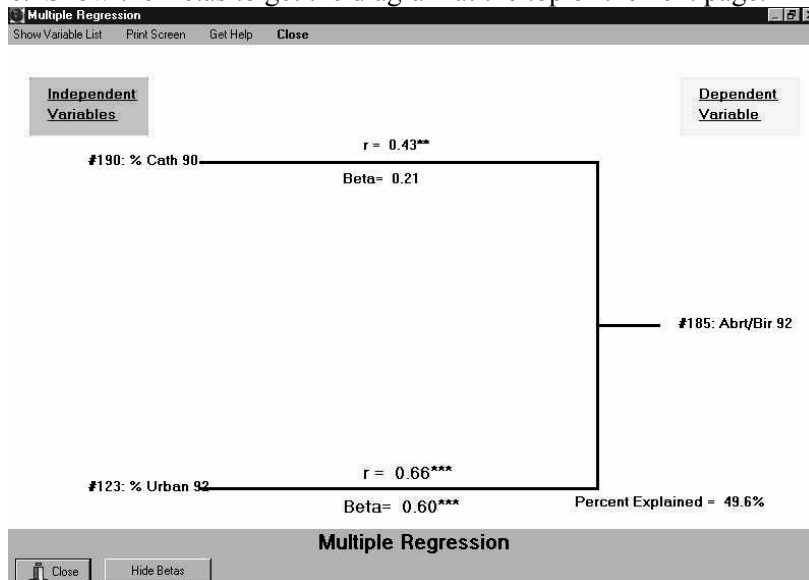
Yet reality is quite different. Rhode Island has the 7th highest abortion rate among the 50 states, Massachusetts has the 6th, and Connecticut the 11th. Alabama is 28th, Tennessee is 32nd, and Arkansas is 37th. And the pattern holds for the rest of the states, as well:

	% Cath 90	Abt/Bir 92
% Cath 90	1.00**** 48	
Abt/Bir 92	0.43** 48	1.00**** 50

The two variables are positively correlated! Not only is the abortion rate not lower in heavily Catholic states, it is actually higher. One would expect the Pope to be pretty steamed by this news.

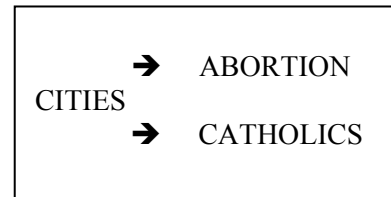
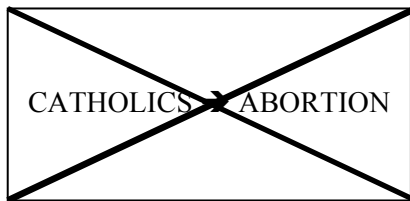
Let's check this, however. We also know that the abortion rate is higher in states with large urban populations, at least partly because abortions are more available there. It is pretty easy for city dwellers to get to an abortion clinic, if they need one. It's a lot harder for rural folks to get to one halfway across their state. All other things being equal, we should expect states in which a higher percentage of the population lives in urban areas to have higher abortion rates.

Set up a regression using **Var 185: ABRT/BIR** as the dependent variable, **Var 190: %CATHOLIC** as the first independent variable, and **Var 123: %URBAN** as the second independent variable. Show the Betas to get the diagram at the top of the next page.



This is very interesting! The regression shows us that the correlation between % CATHOLIC and ABRT/BIR is **spurious**. Holding urbanization constant, there is no relationship between the percentage of the population that is Catholic and the abortion rate. Therefore it is not true that areas

where there are lots of Catholics have lots more abortions. Urban areas have higher abortion rates, and Catholics just happen to concentrate in cities! Here's the logical flow:



Now for the interesting question: Why should the Pope still be upset?

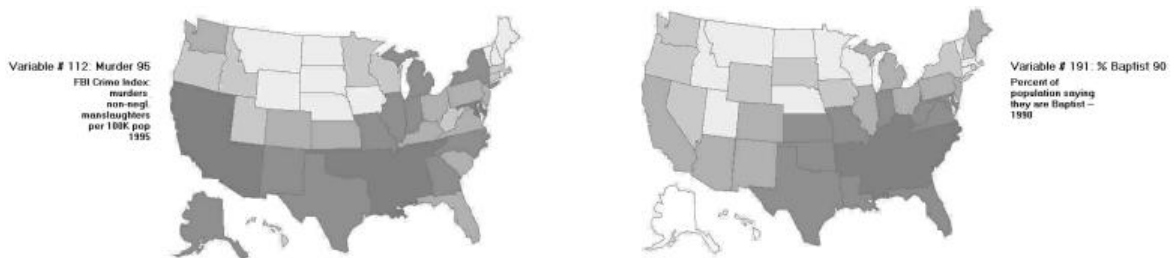
Take a good look at the Beta between % CATHOLIC and the abortion rate. It has no asterisks next to it, which means that there is no predictable relationship between the percentage of Catholics living in a given state and that state's abortion rate. So far, so good.

But if Catholics were actually paying attention to Church teachings, shouldn't the abortion rate be lower in heavily Catholic areas than in areas where Catholics are rare? Shouldn't we see a negative Beta, with asterisks, rather than one that has no asterisks at all? Sure, we now know that heavily Catholic areas are not more likely to have high abortion rates, when one takes urbanism into account. But they are not less likely, either. Someone isn't listening to Church authorities.

Example 2: Do Large Concentrations of Baptists Increase the Murder Rate?

There is a similar connection between the percentage of a state's population that is Baptist and that state's murder rate. The correlation is 0.63***. This is an extremely strong correlation for sociological investigations. Where there are lots of Baptists, there are also lots of murders. Clearly, people in such places are ignoring the 6th Commandment, something that Baptists are supposed to take very seriously.

What's going on? To find out, let's map these two variables. **Choose "Compare Maps"** from the **States** menu. **Type "191"** (% BAPTIST 90) in the first box, then **type "112"** (MURDER 95) in the second. (This data is the closest in time that we have to the 1990 count of various religions.*) You'll get these two maps:



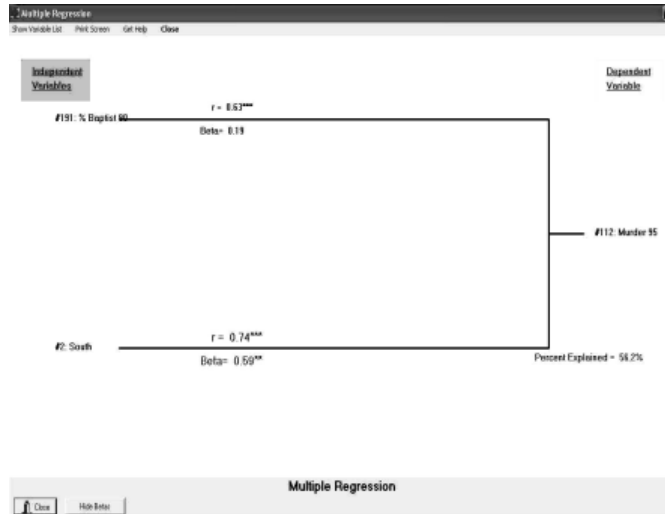
Despite their high correlation, these maps are not identical. As we can see, Baptists are concentrated in the Old South. With some exceptions, murder is concentrated across the southern part of the U.S. (Just for fun, map Variables 220, 221, 222, and 223, to see the long-term murder trend.) Running a correlation table between Variables 112, 191, and 2 shows that % BAPTIST 90 correlates with SOUTH at $r = 0.80^{***}$ and MURDER 95 correlates with the SOUTH at $r = 0.74^{***}$.

* It is also after that count. This is appropriate, as it is less plausible that excessive murders produce Baptists than the other way around.

Multiple Regression (I)

Both of these correlations are higher than the raw correlation between % BAPTIST and MURDER 95. What might be the causal relationship?

Choose “**Regress**” from the **STates** menu. Use “**112**” (MURDER 95) as the dependent variable, use “**191**” (% BAPTIST) and “**2**” (SOUTH) as the independent variables. Click “**OK**” and then “**Show Betas**” to get the following path diagram.



As you can see, the Beta for the SOUTH is significant, while the Beta for % BAPTIST is not. There is no real relationship between the concentration of Baptists in a state and that state’s murder rate. Both Baptists and murder are concentrated in the Old South; it is each one’s independent relationship with the South that creates the spurious correlation between them.

THE ECOLOGICAL FALLACY

There’s one more topic to cover in this chapter before we close. If you’ve paid close attention to my writing, you may know what it is. But you may have missed it as well.

I hope that you noticed how careful I was with my comments about Catholics and the abortion rate. Even though there is a moderate positive correlation between % CATHOLIC and ABRT/BIR, I never said that Catholics were having more abortions than others! Aggregate data are just that: aggregate. They measure things like the percentage of people in a state who are Catholic or Baptist, the rate of infant mortality, the poverty rate, crime rate, and so on. They tell us a lot about what goes on in the various states, but they don’t tell us anything about individuals!

We can, for example, know that 61.9% of Rhode Islanders are Catholic, and that the RI abortion rate stands at 461 for every 1000 live births. We know that New York is 44.3% Catholic, and that its abortion rate is 694 per thousand. But we don’t know who’s getting the abortions!

It may be, for example, that every Catholic in New York and Rhode Island follows Church teachings and stays away from abortion clinics. Given the high abortion rate in these states, that would likely mean that the birth rate for non-Catholics is nearly zero. But – short of contrary evidence – we have to allow for this possibility.

Here’s the point:

Knowing what happens in areas doesn’t tell us what individuals in those areas are doing.

Claiming the contrary is called **the ecological fallacy**. The rate at which something happens in a given area tells us nothing about who in that area is doing whatever it is we are interested in. For example, the rate of Catholicism, police presence, or liberal politics in an area doesn't tell us who the Catholics, cops, or liberals in that area actually are. And it tells us nothing about what they do as individuals. If we claim that it does, we commit the ecological fallacy.

That said, we can still argue that the Pope ought to be upset that abortion rates are no different in areas with lots of Catholics than with few of them. If a state has a large Catholic population, one would expect that population to have some influence over public life and over public policy. Among other things, they vote. They also – like everyone else – influence their neighbors, write letters to the editor, support certain charities and not others. In short, they take part in civic life.

One normally expects that where a large number of citizens share certain core beliefs, those beliefs will affect everyone in their locale. The issue, then, is not that Catholics are or aren't having abortions – we don't have data on that. The issue is that Catholics are not having the kind of influence on abortion policy that Church leaders think they should. Were they having such influence, there would be a lower rate of abortions in heavily Catholic areas – among both Catholics and non-Catholics alike. The above regression analysis tells us that this is not the case. It tells us that there is no difference between Catholic and non-Catholic areas, once urbanism is taken into account. Thus Catholic teaching has no effect – at least its teaching about abortion!

THINGS TO REMEMBER

1. **Multiple regression** sorts out the causal relationships between variables that are intercorrelated. It shows what the correlation between an independent and a dependent variable would be, after removing the influence of other independent variables. The regression analysis provides a **Beta** correlation between each independent variable and the dependent variable.
2. We call this “**controlling**” the influence of the independent variables on each other. Each independent variable's Beta tells us the real relationship between that variable and the dependent variable, after controlling for the influence of all the other independent variables combined.
3. One interprets the Beta much like the Pearson's **r**:
 - a plus sign indicates a positive relationship between the variables; a minus sign indicates a negative one.
 - A large number (either + or -) indicates a strong relationship; a small number indicates a weak one. Use the number of asterisks as a guide
4. **Spurious** correlations are those that regression analysis exposes as false. What appeared to be a correlation vanishes when we take the action of another independent variable into account. (**Sociological Insights**[®] displays real correlations in red and spurious ones in black. Also use the asterisks as a guide.)
5. The “**Percent Explained**=” value is the amount of variation in the dependent variable that can be explained by the independent variables taken together.
6. As a rule of thumb, don't use more than 4 independent variables with the STATES data set. Data sets based on a larger number of cases – such as the 3000 U.S. counties – can handle larger numbers of independent variables. Don't exceed a 15/1 ratio of cases to variables..
7. Don't commit the ecological fallacy by claiming that the **rate** at which things happen in a given place to what specific individuals in that place do.

FIVE: MULTIPLE REGRESSION (II)

This chapter continues our study of multiple regression, especially those cases in which we have more than two independent variables. On the one hand, the task is simple. All we have to do is to enter three or more variables into the right boxes on Sociological Insights' regression form, and the program will calculate "r"s, Betas, and so on for us. But when does one choose to add more than two variables to a regression analysis? And which ones? This takes a bit of thinking.

The easiest way to understand the process is to work through an example. **Start Sociological Insights[®], click the STates menu, choose "Correlate", and type "240", "189", and "133" in the blanks -- the suicide rate in 1991 (except for Nevada, because that state is an **out-lier**), the percentage of the population claiming to have "no religion" a year earlier, and the percentage of the population that moved during the previous five year period. **Click on "OK"** to get this correlation matrix:**

	Suicide -NV	No Relig 90	Moved 85-90
Suicide -NV	1.00*** 49		
No Relig 90	0.51*** 47	1.00*** 48	
Moved 85-90	0.49*** 49	0.64*** 48	1.00*** 50

Leaving Nevada aside, the correlation (in the early 1990s) between the suicide rate and the percent of a state's population that is not religious is a strong +0.51. This means that places with low levels of church membership are also places where there are high suicide rates. Does this mean that not attending church causes people to kill themselves? Not necessarily. We don't know which people in low church-attending states kill themselves. But we do know that places where people don't belong to churches are also places with lots of suicide.

Each of these rates is also strongly correlated with the percentage of the population that moved in the previous five years. Var 189: NO RELIG 90 and Var 133: MOVED 85-90 correlate at +0.64; Var 240: SUICIDE 91 and Var 133: MOVED 85-90 correlate at +0.49. Like variables 132 and 133 or 117 and 118, these three variables are highly **intercorrelated**. Yet they do not all measure the same thing. Which way do the causal relationships run?

One can make various suggestions.

As for residential mobility, it makes a great deal of sense that there would be less suicide in communities where people have lived together a long time and have a lot of social ties. When people have a lot of friends and acquaintances, or are regularly involved with their neighbors, social clubs, civic groups, and so on, they tend not to kill themselves. If they get lonely and depressed (as all people do now and then), they can turn to others for companionship or help. This, in fact, was one of Émile Durkheim's claims in his pioneering book, Suicide: A Study in Sociology.*

* Émile Durkheim, Le Suicide: Étude de Sociologie. Paris: Alcan, 1897. Translated by John A. Spaulding and George Simpson and published by The Free Press (New York) in 1951.

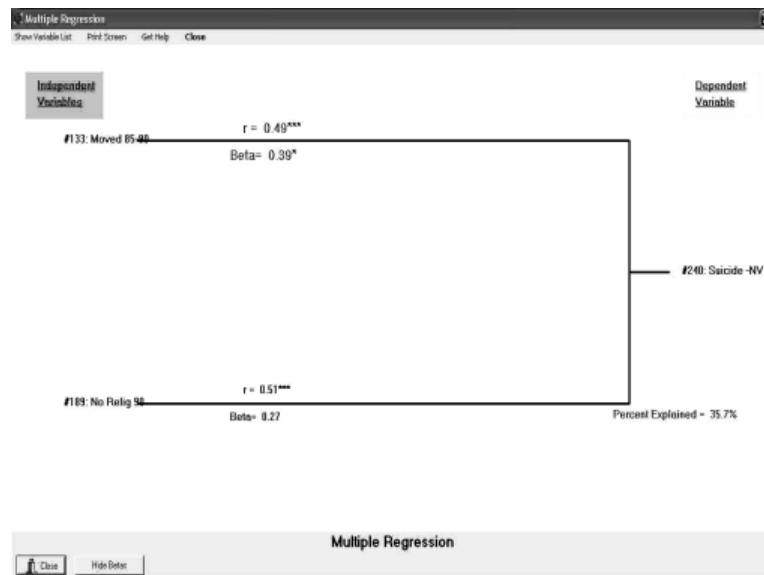
Multiple Regression (II)

Durkheim claimed suicide concentrated in places where social ties were weak. “Egoistic suicide”, he wrote, is the result of a social order that lack sufficient social ties to bind people to each other.

It also makes sense that belonging to a religion prevents suicide – another of Durkheim’s claims.* He argued that at least some religions prevent suicide by regulating people’s morality. Some religions oppose suicide more than do others. Members of those groups will be less likely to kill themselves – and more likely to help others overcome suicidal feelings – because of their religious views. Durkheim said that “anomic suicide” occurs more often where such moral rules are weak or absent.**

Are both of these factors really important? Which is strongest? Which is weaker? Though data on the 50 U.S. states are not the ideal data with which to test these propositions – professional sociologists like to use larger data sets – they do let us see how such testing is done.

Choose “**Regress**” from the **STates** menu, use **Variable 240: SUICIDE -NV** as the dependent variable, **Variable 133: MOVED 85-90** as the first independent variable, and **Variable 189: NO RELIG 90** as the second independent variable. Press the “**Show Betas**” button to get this path diagram:



What does this tell us? First off, we note that the Beta between NO RELIG and SUICIDE is +0.27. It has no asterisks, which indicates that there is no real relationship between them. If all states had the same level of neighborhood stability, the effect of religion on suicide would disappear. The same is not true of the influence of MOVED on SUICIDE, however. Its Beta is still a very robust +0.39. This means that a stable population really does lower the suicide rate. Durkheim was right

* Durkheim’s claim has been subject to considerable sociological critique. See, for example, Lincoln H. Day, “Durkheim on Religion and Suicide : A Demographic Critique”, *Sociology* 21/3: 449-461, 1987; Franz von Poppel and Lincoln H. Day, “A Test of Durkheim's Theory of Suicide -- Without Committing the 'Ecological Fallacy'”, *American Sociological Review*, 61: 500-517, 1996; and Miles Simpson, “Comment: Suicide and Religion: Did Durkheim Commit the Ecological Fallacy, or Did van Poppel and Day Combine Apples and Oranges”, *American Sociological Review*, 63: 895, 1998

** I have posted a chart showing the relationship between Durkheim’s various kinds of suicide (of which “egoistic” and “anomic” are but two) at <http://www.socialdata.info/DurkheimChart.htm>.

about the importance of social ties. He was wrong, however, about moral ties, because the presence or absence of religion has no effect on the suicide rate. If it had an effect, we would see that both factors are significant.

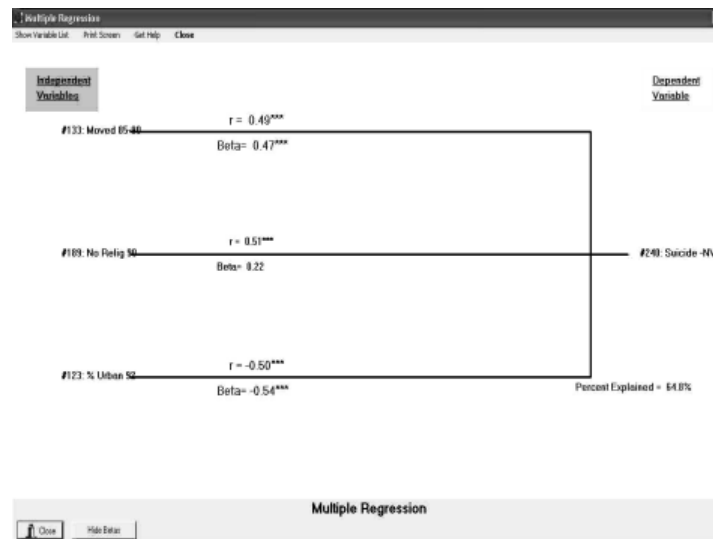
Well, that's not quite true. In this diagram, religion and population stability together explain nearly 36% of the variation in suicide rates. In the last chapter, however, we saw that MOVED by itself explained only 24% of the variation in suicide rates. Adding NO RELIG explains another 12%. This tells us that NO RELIG makes more difference than did the SEX RATIO – which added nothing – even though the Beta between NO RELIG and SUICIDE –NV is too small to indicate any independent influence.

What do we do in such a situation? The lack of asterisks says there is no real relationship, but the change in the percent explained says that there is something going on. This question has two answers.

First, we stand by the lack of asterisks. Though there may be a relationship between religion and suicide, after social stability is held constant, it is too weak to be counted on. The correlation between NO RELIG and SUICIDE is spurious. At the same time, there is something going on here which deserves our attention. Religion's impact on the percent explained tells us that MOVED is not the whole picture. This leads us to investigate other factors that might be involved.

MORE THAN TWO INDEPENDENT VARIABLES

Sociologists have long known that city life affects people's social ties. Yes, people move around a lot in cities, but some cities are much more stable residentially than are others. What happens if we add the state-by-state variation in urban life to our analysis. Set up a regression using “240” (Var 240: SUICIDE -NV) as the dependent variable and “133” (Var 133: MOVED 85-90), “189” (Var 189: NO RELIG 90), and “123” (Var 123: % URBAN 92) as three independent variables. Add the Betas to get the path diagram:



Here, both MOVED and % URBAN have real and independent relationships with the suicide rate. NO RELIG (again) does not – and its Beta has shrunk to +0.22. The three variables together explain nearly 65% of the state-to-state variation in the suicide rate. That's considerably more than did MOVED and NO RELIG, or than MOVED alone. This tells us that %URBAN is a major factor influencing the suicide rate, though a negative one.

It is important to realize that each of these independent variables **controls for** the others. This lets **Betas** indicate not just the independent influence of each, but also their relative strength. With a Beta of -0.54^{***} , city life is the strongest predictor of the suicide rate among these variables. It tells us that states with large urban populations have lower suicide rates – after controlling for those states’ level of residential mobility and degree of (or lack of) religion. The Beta of 0.47^{***} isolates the effect of residential mobility on suicide: it tells us that states with high levels of residential mobility – in cities or outside of them – tend to have higher level of suicide than others. Again, religion is not independently important.

Normally, we would remove religion from our analysis at this point, since it has failed two regression tests. However, we are testing Durkheim’s hypothesis that religion prevents suicide independently of social ties, so we won’t throw it out just yet.

In this regression analysis, MOVED and % URBAN are **predictor variables**. If we know how much of a state’s population moved during the last year and how much of that population lives in urban areas, we can predict that state’s suicide rate. States with lots of residential mobility have higher suicide rates, and states with lots of city dwellers have lower suicide rates. These two variables operate independently of one another. On the other hand, NO RELIG is not a predictor variable. Knowing the percent of a state’s population that does not belong to a religious organization does not let us predict that state’s suicide rate – at least not after controlling for our two other predictor variables: residential mobility and urbanization.

We could continue this process, though explaining 65% of the suicide rate is doing quite well by sociological standards. By testing one variable after another, one can gradually identify the variables that influence that rate and can eliminate those that don’t. Gradually, one pushes the “**Percent Explained** =” value as high as one can.

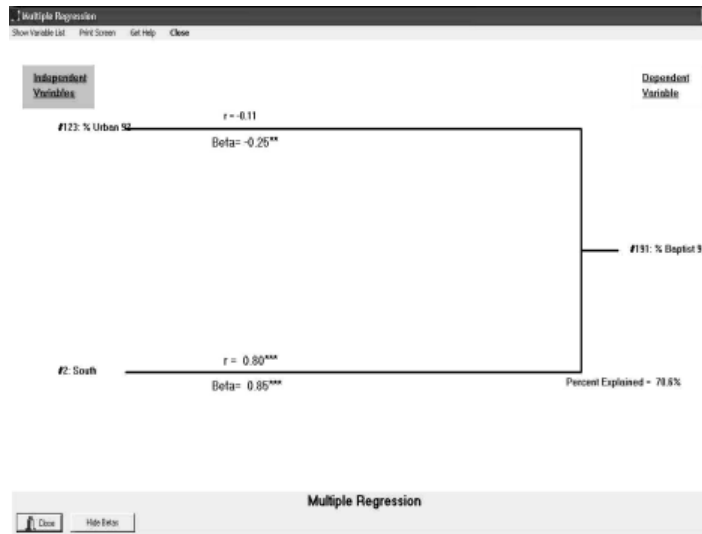
SPURIOUS NON-CORRELATION

We still have one concept to go. Sometimes, we find that two variables are not correlated, but only because they are both related to a third variable that suppresses the interaction between them. For example, **map** the three variables **191: % BAPTIST**, **123: % URBAN**, and **2: SOUTH**. The first and third of these maps look very similar. The second appears not to be correlated with either. And that is what a **correlation matrix** tells us:

	% Baptist 90	% Urban 92	South
% Baptist 90	1.00 ^{***} 48		
% Urban 92	-0.11 48	1.00 ^{***} 50	
South	0.80^{***} 46	0.18 46	1.00 ^{***} 46

There is, in fact, an overwhelming connection between Baptists and the states of the Old South. There is also little relationship between the Baptists and urbanism and between the South and urbanism – Pearson’s “r” is -0.11 for the former and $+0.18$ for the latter, each with no stars. :

Now run a regression analysis, using “**191**” (% BAPTIST) as the dependent variable and “**123**” (% URBAN) and “**2**” (SOUTH) as the independent variables. **Click “Show Betas”** to get the path diagram at the top of the next page.



Surprised? When we remove the influence of region, urbanism is revealed to be negatively correlated with the percentage of the population that is Baptist. What is going on here?

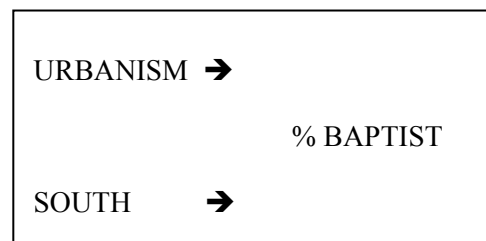
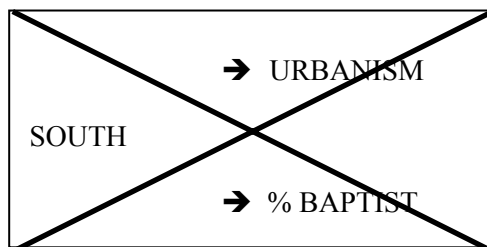
Let's review what we know. We know that states in various parts of the country are differently urbanized. We know that Baptists live in the disproportionately in the South. It just so happens that the more urbanized Southern states have more Baptists than do the more rural Southern states. Exactly the opposite is true in the Northeast, Midwest, and West. There, the Baptists disproportionately live in rural states, not urban ones.*

When we put these patterns together, we find that outside the South, there is a clear negative relationship between urbanism and percent Baptist. This relationship is masked by the fact that inside the South there is a clear positive one. These patterns cancel each other out, until we control for Variable 2: SOUTH. Then the pattern outside the South emerges.

Spurious relationships are those relationships that disappear when we use a regression to control for the influence of other variables.

Hidden spurious relationships are those relationships that appear when we control for other variables.

We have just encountered a common phenomenon in quantitative sociology. More often than we might think, variables that appear to be uncorrelated are actually correlated. Sociologists need to be on the lookout for such **hidden spurious** relationships. They can hide relationships that are important to see. In this case, the box on the right shows the true relationship between these three variables. The box on the left does not.

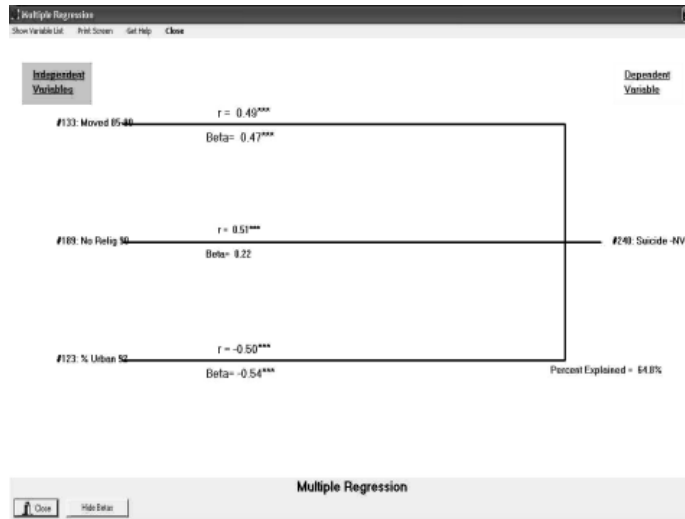


* We don't know, of course, whether individual Baptists live in cities in the urban states and rural areas elsewhere; they may live in cities in the latter and on farms in the former. Regression analyses work with aggregate, not individual, data.

Multiple Regression (II)

Now let's return to our analysis of suicide rates, because there is an important hidden spurious relationship alongside the spurious relationships that we have already viewed.

You remember that we added Variable 123: % URBAN to our previous analysis because Variable 189: NO RELIG contributed significantly to the “Percent Explained”, even though it was not itself a significant predictor of suicide. Adding urbanism alongside residential mobility and lack of religion explained nearly 65% of the variation in the state-to-state suicide rate.

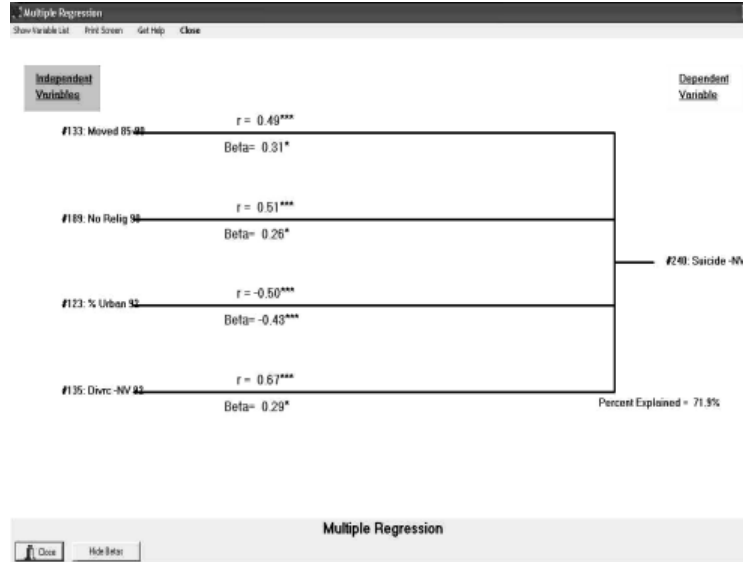


Were we to drop NO RELIG from this regression, however, we would drop the “Percent Explained” to a bit over 50%. Even though a state’s relative lack of religion does not let us predict that state’s suicide rate, NO RELIG appears to have some residual influence.

What happens if we add the divorce rate to this mix? After all, the Pearson’s correlation coefficient (“r”) between Variable 135: DIVORCE -NV and Variable 240: SUICIDE -NV is 0.67*** – a very strong correlation. Like residential mobility, a high divorce rate is a sign of lack of social ties. It is not, however the same lack of social ties, so we do not have to worry about entering the same thing twice. States with low church membership rates have moderately higher divorce rates than do other states. Perhaps DIVORCE will finally shove NO RELIG out of the picture.*

Run a regression, using “240” as the dependent variable and “133”, “189”, “123”, and “135” as the independent variables. Click “OK”, then click “Show Betas” to get the diagram on the top of the next page.

* We use Variable 135: DIVORCE -NV rather than Variable 134: Divorce 92, because Nevada is as much an outlier for its divorce rate as it is for marriage and for suicide.



Clearly, this time NO RELIG does not disappear! In fact, the percentage of a state's population that is not religious is a significant predictor of that state's suicide rate, after controlling for residential mobility, urbanization, and divorce. None of our four independent variables are spurious. All influence the divorce rate independently of one another.

What else can we learn from this path diagram? Here's a quick summary, in order of the variables' importance:

1. States with large percentages of urban residents have lower suicide rates than do states with lower percentages of city dwellers – after controlling for residential mobility, religion, and divorce. This is the strongest predictor of the suicide rate, and it is negative.
2. States with large amounts of residential mobility have higher suicide rates, after controlling for the other three factors.
3. States with high divorce rates have higher suicide rates. Marital ties seem to operate independently of neighborhood ties and church ties in preventing suicide.
4. States with low church membership have higher suicide rates. Church ties thus seem to operate independently of neighborhood ties and marriage ties in preventing suicide.
5. In all, these four factors explain nearly 72% of the state-by-state variation in suicide rates. (This is a huge amount for sociology; we normally have to settle for much less.)*

What does this result have to say to Émile Durkheim? Remember that he posited two separate social influences on the suicide rate. One was the absence of sustaining social ties. The other was the absence of moral regulation. I think we can make a plausible case that we are seeing both at work here – and in precisely the direction that Durkheim predicted.

Three of the above four factors mark the presence or absence of social ties. First, states with highly urban populations have less suicide, after controlling for the other factors, because cit-

* This analysis works for 45 of the 50 U.S. states. Besides leaving out Nevada, we had no church membership figures for Alaska and Hawaii and so had to leave them out of the analysis. We had no divorce figures for Indiana and Louisiana, so we had to leave them out as well. Given this, we would be well advised to check our conclusions on a fuller data set. But the procedure for doing so ought to be clear.

ies give people a great many more possible social ties than do suburban and rural areas. Not everyone will take advantage of these ties, but those who want them can find them. There are just lots more people around. This produces a lower suicide rate. Second, highly mobile populations have fewer social ties, which (after controlling for the other factors) produces a higher suicide rate. Third, divorce is a clear breakdown of a very important social tie; places where there is more divorce are shown to be places where there is more suicide – again after controlling for the other factors.

Our latest regression, however, shows that religion is in fact also a predictor of suicide – but only after we have cleared several different kinds of social ties out of the way. Urban ties, neighborhood ties, and marriage ties are all different kinds of social ties. Once we control for these ties, we still have something left over. This either indicates the existence of a specifically religious tie – belonging to a church, synagogue, or mosque is a form of social tie, after all. Or it indicates the existence of the kind of moral regulation whose absence, Durkheim claimed, resulted in anomic suicide. I suggest that the latter is more plausible, with a two-step argument:

- First, the degree of religious involvement of a state's population appeared to have no relationship to its suicide rate until after we had cleared several different forms of social ties out of the way. This, by itself, indicates that religion is significantly different from each of these forms of social connection.
- Second, the social ties that one finds in religious groups are not that much different than the social ties that one finds in cities and in neighborhoods. Like these ties, getting involved in church requires time and a sense of social commitment. Long-term residents are more willing to commit themselves to this effort than are people who are less invested in their communities (or who do lack the convenient access that city life brings).

If religion's social ties are pretty much like these other social ties, yet religion has independent influence, that influence must be moral. We have not proved Durkheim's point, but we have lent it some credibility.

How far can we take this? How many variables can we include in a regression analysis? It all depends on how many geographical areas we are comparing. With data from the 50 states, we will usually limit ourselves to four. Sociological Insights[®] has room for 5 independent variables, but this is one more than is really legitimate with this data. As a rule of thumb, don't use more than 3 or 4 independent variables with the States data. We may break this rule occasionally, but we must be very careful about any conclusions we draw when doing so.*

One final warning. We also want to make sure that our independent variables are not just measuring the same thing. If they are, then the mathematics that makes regression analysis possible goes all weird on us – and gives us untrustworthy results. For example, you should never enter Var 132: STABLE 85-90 and Var 133: MOVED 85-90 as two independent variables in the same analysis. They are identical to each other – one capturing the people who didn't move between 1985 and 1990 and the other capturing the people who did. Such **multicollinearity** makes hash of regression procedures.

* Technically, the number of independent variables we can use and still get an accurate regression analysis depends on the number of cases in our data set (**n**), and in this case, **n** = 50 (states). If we were analyzing data on the 3000+ US counties, we could use a larger number of independent variables – as many as 200 (though no one would ever use that many). The general rule is that one should use no more than $n/15+1$ independent variables in any regression analysis. (For details, see Rae Newton & Kjell Rudestam, Your Statistical Consultant, Sage Publications, 1999, pp 250-252.)

Note that your computer will give you results in such cases – because it can't tell the difference between multicollinear and non-multicollinear variables. You'll get path diagrams, betas, percents explained, and so on in either case. You just won't be able to trust such figures, because they'll be mathematically meaningless. And Sociological Insights[®] won't inform you of that fact!

Remember: computers are very fast, but very dumb. Unlike people, they do exactly what you tell them to – even (as we'll see a few chapters hence) give you an average for a sample's sex or race. You always have to be careful to give good instructions.

OTHER KINDS OF REGRESSION ANALYSIS

The kind of regression that we have been using here is **ordinary least-squares (OLS) multiple regression**—a name based on the mathematical formula that drives the process. **Stepwise regression** and **hierarchical regression** are forms of OLS regression. They differ only in the order in which they sort through a series of independent variables to see which ones most influence the dependent one. All these types of regression require interval/ratio data for the dependent variable, as well as for almost all of the independent variables. (One can occasionally use **dummy variables** to handle a categorical independent variable or two, though doing so is beyond the scope of this book.)

Logistic regression is designed for cases where the dependent variable is categorical. Let's say that you're trying to find out what differentiates states that have active capital punishment programs from states that don't. Logistic regression would let you sort states into two categories, those with such programs and those without. You could then propose a set of independent variables, each of which records interval/ratio data. A logistic routine would let you locate the (presumably) causal factors.

Like dummy variables, logistic regression is beyond the scope of this book, and also beyond the scope of Sociological Insights' database. You don't need to worry about whether your data are of the right form for OLS regression; they are.

You now have all the concepts you need to analyze multiple regressions. You may, however, run into a few more terms in other books. Here are three:

- The "Percent explained" is also called the **combined effect** of these variables – the total effect that two (or more) variables have on a third. This is also sometimes called the **explained variance**.
- A variable's Beta is also called the **net effect** of that independent variable when the other independent variables are **held constant**.
- **Holding a variable constant**, of course, means imagining that this variable does not vary from state to state. (Fortunately, the computer does this imagining for you mathematically; you just need to push the buttons.) This is also called **controlling for** that variable's influence.

THINGS TO REMEMBER

1. Be sure to think through the logic of your regression analyses. One doesn't just throw things in willy-nilly. One uses regression to test sociological hypotheses, not to conduct fishing expeditions.
2. **Multiple regression** sorts out the causal relationships between variables that are intercorrelated. It shows what the correlation between an independent and a dependent variable would be, after removing the influence of other independent variables. The regression analysis provides a **Beta** correlation between each independent variable and the dependent variable. This Beta tells us the real relationship between that variable and the dependent variable, after controlling for the influence of all the other independent variables combined.
3. One interprets the Beta much like the Pearson's **r**:
 - a plus sign indicates a positive relationship between the variables; a minus sign indicates a negative one.
 - A large number (either + or -) indicates a strong relationship; a small number indicates a weak one. Use the number of asterisks as a guide
4. **Spurious** correlations are those that regression analysis exposes as false. What appeared to be a correlation vanishes when we take the action of another independent variable into account. (*Sociological Insights*[®] displays real correlations in red and spurious ones in black. Also use the asterisks as a guide.)
5. Sometimes, relationships are **hidden spurious** relationships. In such cases, an original lack of correlation is revealed to be a real relationship – once the masking effect of other independent variables is removed.
6. The “**Percent Explained**=” value is the amount of variation in the dependent variable that can be explained by the independent variables taken together.
7. As a rule of thumb, don't use more than 4 independent variables with the STATES data set. Data sets based on a larger number of cases – such as the 3000 U.S. counties – can handle larger numbers of independent variables. A 15/1 ratio of cases to variables is about the most one can use.
8. Don't use two independent variables that measure essentially the same thing, as it makes the math underlying OLS regression fail. Statisticians call this **multicollinearity**, and avoid it.
9. Don't commit the ecological fallacy by arguing from the **rate** at which things happen in a given place to what specific individuals in that place do.
10. Like correlations, **ordinary least-squares (OLS) multiple regressions** require **interval/ratio data**. If one's dependent variable is categorical, one can use **logistic regression** – a technique not covered in this book. If one of the independent variables happens to be categorical, one can sometimes use **dummy variables** in OLS routines. This, too, is beyond the scope of this book. (Don't worry; *Sociological Insights*[®] doesn't have any such variables in its database.)

SIX: DISTRIBUTIONS AND CROSS-TABULATIONS

The previous chapters have explored **aggregate data** based on the 50 U.S. states. As I pointed out previously, there are other kinds of aggregate data – including some based on groups and social categories (e.g.: race, gender) rather than on geographic units. But the statistical routines for examining them are the same.

The next several chapters will explore a different kind of data: **survey data**. These data come from interviews with individuals. Yes, sociology studies groups and group behavior. But one of the best ways of finding out what is happening in groups is to talk to individuals. If we ask a lot of people the same questions, we get a pretty good picture of what's going on in society at a whole.

The survey we are going to use is called the **General Social Survey (GSS)**. Nearly every year from 1972 to 1992, and every 2nd year thereafter, the National Opinion Research Center (NORC) interviews a large number of American adults. Questions cover everything from their socioeconomic status to their family life and to their attitudes on social issues. They even ask how happy people are. The early surveys covered about 1,400 people. The more recent ones have covered about twice that. Though these are a tiny fraction of the 250 million U.S. residents, the way the survey is conducted lets us generalize from this small **sample** to the population as a whole. This is because the sample is **random**. This means that every English-speaking, U.S. resident over 18 who is not in the military, a hospital, jail or mental institution has an equal chance of being interviewed in any given year. By the mathematics of probability, 1,400 respondents can pretty well represent the United States.* It is true that small minorities – for instance, Albanian-Americans – are not accurately depicted. But major groups stand out, both in their similarities and their differences. And the GSS is carefully randomized, making it probably the best such data set available.

The data we will look at come from the 2000 General Social Survey. I have selected nearly 250 variables to consider. That year, NORC polled nearly 2,817 persons. On the other hand, it did not pose every question to every respondent. Thus sometimes we will have more than 1,500 answers to a question, and sometimes fewer. Don't worry. The software will tell us how significant the answers are.

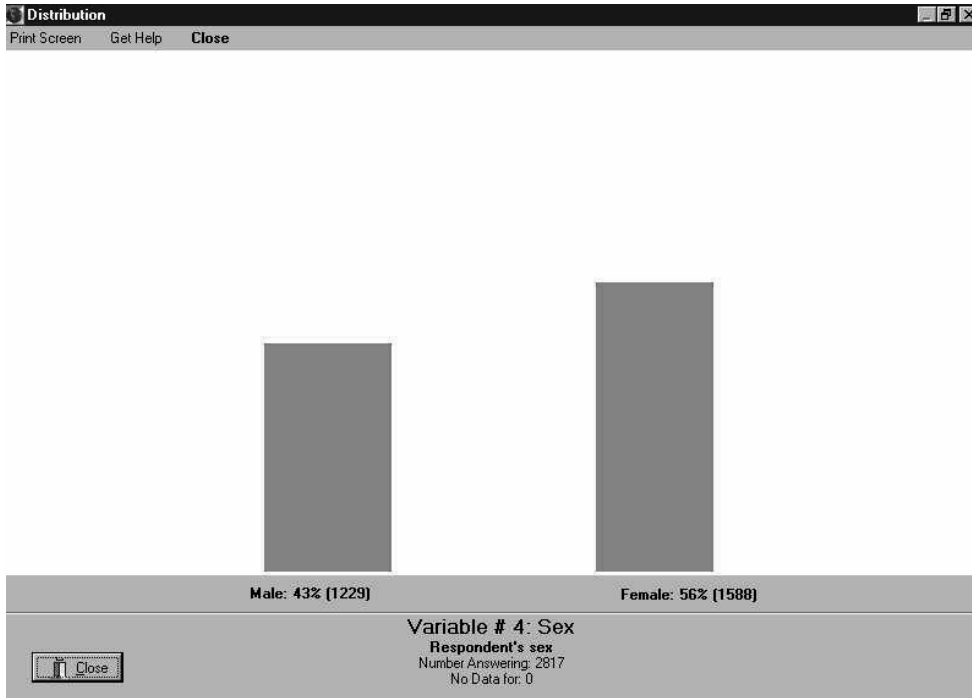
(The Sociological Insights[®] CD also has data from the 1994 GSS. That year NORC also interviewed nearly 3,000 Americans. This data set only contains 113 variables. That's not because the GSS asked fewer questions; I just didn't select as many questions to use.)

DISTRIBUTIONS

Start Sociological Insights[®]. Click on the **SURvey** menu, then choose **“Distribute”** to show a variable's **distribution**. This is a summary of the number of people who give each possible answer to a survey question. **Type “4”** (Var 4: SEX), then **click on “OK”**. This displays a bar chart of the

* As regular statistics books will tell you, random sampling works because it allows us to predict the distribution of scores for a very large population. The **central limit theorem** tells us that the distribution of random sample means approaches a particular mathematical curve as sample size increases – regardless of the actual distribution of the population. (The curve is called the “normal” curve.) It turns out that, 19 times out of 20, the mean score of a random sample of 1,400 persons will be within a couple of percentage points of the mean score of any size population. The recent GSS use of nearly 3,000 people gets us even closer – except that the newer surveys don't ask every question of every person. GSS answers still constitute the most accurate depictions available of the adult, English-speaking, non-institutionalized American population.

number who answered "male" and the number who answered "female" when asked for their gender. You should see the following:



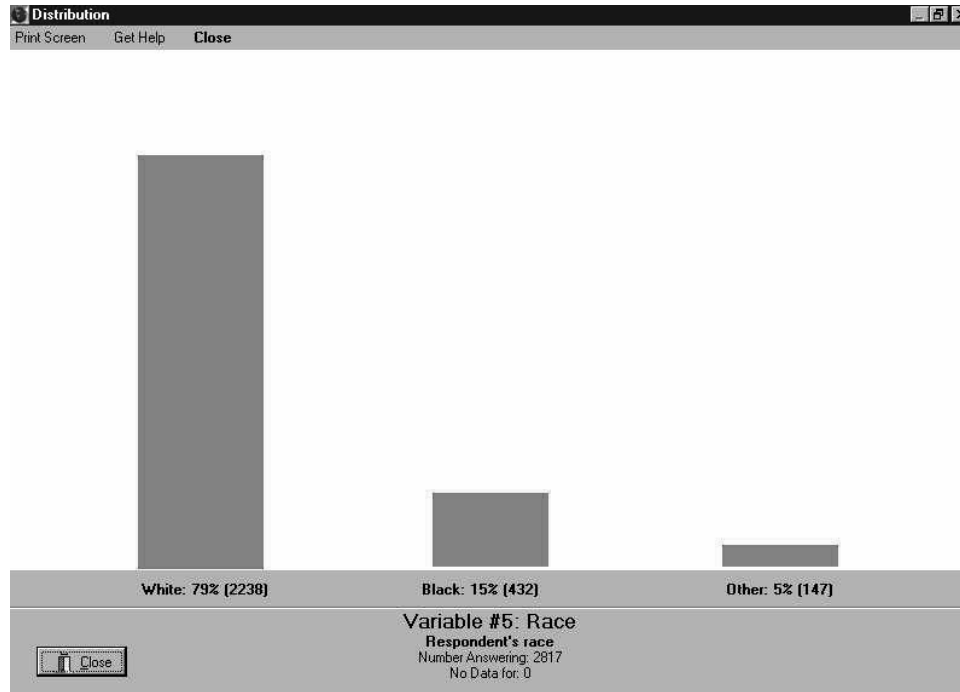
Several parts of this display are noteworthy. The most important, of course, are the two bars, whose heights reflect the gender ratio of our sample. Right under the bars are the actual counts and percentages: from a total of 2817 respondents, 1588 (56%) are women and 1229 (43%) are men. (These percentages don't add up to 100% because of rounding.) This ratio seems surprising. Aren't men and women each born in about the same numbers? Why are 56% of the respondents women?

There are three important reasons for this skewed distribution. First, the GSS surveys adults. Men tend to die younger than women, so there are more females than males over the age of 18. Second, the GSS does not survey people in jail or in the military – who are overwhelmingly men. Third, a sample this size is accurate to within 2 percentage points. This is pretty good, but it means that women could actually be as low as 54% of the adult, non-institutionalized population. Female longevity and low jail rates explain the rest of the discrepancy.

Let's check out the other important figures. At the bottom center we see the number of people who answered this question – 2817. We also see the number for whom we have no data: either they were not asked the question, they refused to answer, or they said that they did not know. In this case, everyone answered. But sometimes they do not. It helps to know this when we are trying to interpret the graphs.

By the way: this variable is a good example of categorical data. At least as far as the General Social Survey is concerned, one must be either "male" or "female". These unranked categories are the only options.

Close this distribution chart, then **type "5"** (Var 5: RACE), then **click "OK"**. The chart on the top of the next page gives us a breakdown by race.



Of the 2817 respondents, 79% are White, 15% are Black, and 5% are Other. Data elsewhere in the data set tell us that half of these “Others” say they are Hispanic or Latino/a; most of the rest are likely Asian.

In fact, 6% of Whites and 2% of Blacks also say that they are Hispanic or Latino/a. This is because “Hispanic” is an ethnic, not a racial, category. The term “Hispanic” (invented by the Census Bureau) lumps Spanish Europeans, Guatemalan Indians and Afro-Cubans together, along with many others. The [Statistical Abstract of the United States](#), which has listed Hispanics separately for only the last two decades, notes that “Hispanics can be of any race.” Some “Hispanics” consider themselves White, others consider themselves Black, while still others consider themselves to be a separate racial group.

Still, the General Social Survey probably underrepresents people of Latin American descent. Remember that the GSS only interviews English speakers. Though most Hispanics in the U.S. are English speakers, some are not. The survey skips them, just as it skips the men in jail mentioned above.

Now you get to explore some of these distributions on your own. **Close** this graph, then **type “2”** (Var 2: AGE) and **click “OK”** to see the distribution by **age**. Then look at:

- Marital status (**Variable 22: MARITAL**)
- Religious affiliation (**Variable 96: RELIGION**)
- The age of respondent when her/his first child was born (**Variable 31: AGE 1ST KID**)
- Who people voted for in 1996 (**Variable 51: VOTE WHO 96**)
- How often respondents spend an evening with their relatives (**Variable 148: VISIT KIN**)*
- How often they spend an evening in a bar (**Variable 151: BAR NIGHTS**)*

* Their answers to these two questions are grouped into several ordinal categories: “1-Sev/Wk” means once or more per week; “1-Sev/Mo” means at least once a month but not as often as once a week; “1-Sev/Yr” means at least once a year, but not as often as once a month”; “Never” (of course) means never.

Any surprises?

(Of this set, variables 22, 96, and 51 are categorical data, and variables 148, 151, and 31 are ordinal data. The difference is that the latter can be ranked, and the former cannot. In the original GSS files, though, variable 31 was interval/ratio data – the raw age at which one’s first child was born. I have transformed it to ordinal data to fit the analyses that Sociological Insights[®] uses for social surveys.)

CROSS-TABULATIONS

Distributions are interesting, but they only go so far. It's nice to know what percentage of Americans consider themselves happy (Variable 113: HAPPY?), for example, (31% are "very" happy, 57% are "pretty" happy, 10% are "not too" happy), or how many are satisfied with their marriages (Variable 114: HAP MAR?). But it's more useful to know who is happy and why they are.

We can get this information from the GSS by comparing variables with each other. By looking at gender and happiness simultaneously, for example, we can tell how many of the women are "very" happy, how many are "pretty" happy, and how many are "not too" happy. And we can compare them to the men. (Do you think there is any difference?) We call this technique "**cross-tabulation**," because it makes a table of all possible combinations of answers and tells us how many respondents' answers fit each of the table's cells.

Let’s check up on happiness. **Hit <ESC>** until you are at the main Sociological Insights[®] screen. **Open the SURvey menu and choose “Cross-Tabulate”**. **Type “4”** (Var 4: SEX) in the blank for the independent (column) variable, then **type “113”** (Var 113: HAPPY) in the blank for the dependent (row) variable. :Leave the third box blank; we won’t use that until chapter 8. **Click “OK”** to see the table:

		Independent Variable: Sex (Respondent's sex)						ROW TOTALS
		Male	Female					
Dependent Var.: Happy? (In general - are you very happy - pretty happy - or not too)	Very	31.2% (379)	32.1% (502)					31.7% (881)
	Pretty	58.1% (705)	57.5% (898)					57.7% (1603)
	Not Too	10.7% (130)	10.4% (163)					10.6% (293)
COLUMN TOTALS		100% (1214)	100% (1563)					

Influence of Var #4 (Sex) on Var #113 (Happy?)
2771 observations
Probability = 10.24% (p = .976)

The two variables are written at the left and the top: HAPPY and SEX. The top of the table contains the column labels ("Male", "Female"), marking the options for that categorical variable. The

row labels are on the left (“Very”, “Pretty”, “Not Too”), marking the options for that ordinal variable. (One can rank “very”, “pretty”, and “not too”, though one does not know the exact distance between them.) At the intersection of each row and column, you find two figures: a percentage in boldface and a number in parentheses. The numbers are the numbers of people falling in each cell. For example, 379 men said that they were very happy, as did 502 women. 130 men and 163 women said they were not too happy, etc.

Look at the percents. Unlike the numbers in parentheses, these are not the percentage of the total respondents who fall in each cell. Instead, they are the percentages of people in each column who give a particular answer. Thus, 31.2% of the men said they are “Very Happy”, 58.1% said they are “Pretty Happy”, and 10.7% said they are “Not Too Happy.” Among the women, the percentages are 32.1%, 57.5%, and 10.4%, respectively. This allows us to compare the columns to see how alike they are. We can then tell whether men or women report being happier, as a general rule.

(If you forget how to read these figures, left-click your mouse over any of the cells in the table, and move the mouse just a little bit. If you do this in the cell for very happy men, for example, Sociological Insights[®] will tell you that “31.2% of the 1214 people who answer Male also answer Very.”)

Compare column percents to find out whether people in the categories measured by the column variable differ on the attitudes or traits measured by the row variable

Why do we use percentages? We need them because the GSS interviews different numbers of men and women. If we didn’t compare percentages, we’d be comparing raw numbers and we might think that women were happier than men. After all, more women than men report being “very happy”. But more women than men also report being “pretty happy” and “not too happy” – just because there are more women in our sample! Raw numbers don’t really tell us which sex claims to be happier. But the percentages tell us that men and women have pretty much the same level of happiness, because the columns aren’t too different.

But how different is “too different”? How do we quantify a subjective statement like that? Look near the bottom of the screen. See where it says “Probability = 10-24%”? This means that there is a probability of between 10% and 24% that the small difference we see between these columns in the GSS sample reflects a real difference between men and women in the American population. That’s less than the probability of getting a heads when you flip a coin. This tells us that it is extremely unlikely that the small difference we see between these columns reflects a real difference between American men’s and women’s levels of happiness. Instead, the difference we see is due to **sampling error** – it is the result of the particular set of people that happened to be interviewed for the 2000 General Social Survey.

Though you can’t see it doing so, the program has figured a “**chi-square**,” one of many different **tests of significance**. Like the significance test we used with Pearson’s correlations on aggregate (state-by-state) variables, the chi-square tells us whether the difference between the table’s columns reflects real differences in the American population, or is just due to random chance. The chi-square compares the actual distribution of people across the cells with the distribution that they would have, were there no relationship between the variables. Rather than read the chi-square directly, the program presents a “**Probability**” figure to tell us the probability that the difference in the columns is not due to chance.

If “Probability > 95%”, there is at least a 95% chance that the difference between the columns reflects real differences in the total population. If it is greater than 99%, there is at least a 99% chance that the difference is significant. Social scientists don’t like bad odds, so if the

"Probability" does not reach 95%, we will assume that the difference is random and does not reflect real differences in the underlying data.

(Sometimes, statisticians use "p" instead of "Probability". p is just the opposite of the probability; it measures the odds that the difference is due to chance. For example, when Probability > 97.5%, $p < 0.025$. Sociological Insights[®] gives us both to keep everyone happy.*)

So where does the chance come from? Remember that the GSS only interviews a small number of Americans – 2817 of them in the year 2000. Every non-institutionalized English-speaking American adult has an equal chance of getting in the pool, but it is still quite likely that the pool is a bit different than the population as a whole. Had different people been interviewed, they might have given different answers. As samples get larger, however, the likelihood of these differences being large is greatly reduced. The underlying mathematics is complex, but the short version is that "Probability" calculation tells us how likely it is that the differences between the columns in our sample stem from real differences in the American population. (And "p" tells us the likelihood that the difference is due to sampling error.)

"Probability" must be at least 95% (and $p \leq 0.05$) for us to conclude that the columns are really different. If the figure is less than 95%, then there is **no difference** between the columns; the apparent difference is due to the ways that our sample differs from the population at large.

Remember: only probabilities of at least 95% are good enough for us. A 20-24% probability is not enough to say that the small difference we see here between men's and women's happiness reflects Americans' happiness as a whole. The sampling error is just too large.

(If you hover your mouse over the "Probability" statement, Sociological Insights[®] will tell you whether the difference between the columns is large enough for you to conclude that there is at least a 95% chance that it reflects a difference in the whole population. Try it!)

* Actually, the "p" figure is also a probability measure. It gives you the probability of making what is called a **Type I Error**. This is the error of saying that there is a difference between two things based on a sample when there is actually no difference between them in the population as a whole.

In statisticians' peculiar language, Type I Error is "the probability of incorrectly rejecting the null hypothesis." (The "null hypothesis" is the hypothesis that there is no difference between the columns of a cross-table, between experimental subjects and a control group, etc.). " $p < .05$ " means that there is less than five chances in 100 – one chance in 20 – that we will claim a difference based on our sample where none exists in the world at large.

A **Type II Error** is the probability that one will claim that there is no difference in the larger population when there actually is one. Reducing the risk of Type I Error – i.e., by requiring "p" to be as small as possible – increases the risk of Type II Error. There's no way around this: you either minimize your chance of falsely claiming a difference between the columns or you minimize your chance of missing a real difference between them. Because of this, each scholarly discipline sets its own standards for accuracy.

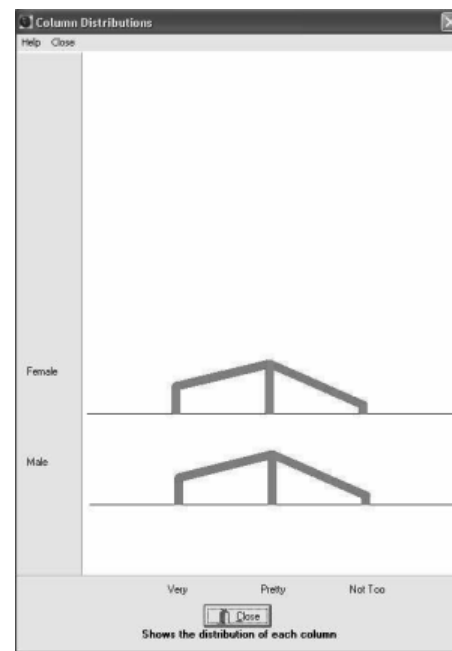
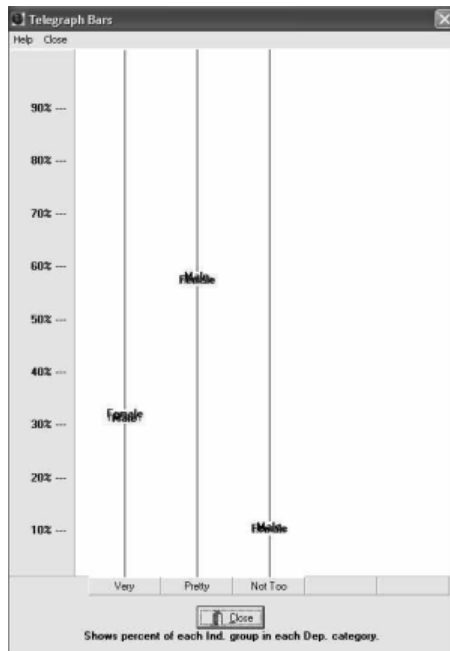
Medical researchers usually set $p < .001$ (or less), because they want to make *really* sure that the difference they see between their experimental group and the control group is not due to sampling error. Not only are drugs expensive to develop – making it only worth developing those that show great promise. Drugs also have side-effects that need to be avoided. Medical researchers would rather miss a few cures than do harm. Thus their preference for 1 in 1000 odds.

Sociology is not such a dangerous matter, so we accept a much more relaxed 1 chance in 20 of Type I Error.

There are a couple of other things to note about the cross-tabulation screen before we move on. First, just below the table title at the bottom of the screen, we see that the table includes responses from 2777 interviewees. Apparently, not all 2817 respondents answered both questions. The fewer respondents, the larger the difference between the columns has to be to reach the 95% threshold. That’s why survey researchers try to get as many respondents as they can afford.

The remaining numbers in the table are what we call **marginals**. Those at the ends of the rows are the total number of both men and women giving each answer – along with the percent of the total number doing so. These give us the distribution of the population as a whole. The numbers at the bottoms of the columns are the total number of men and women, respectively. Because Sociological Insights[®] always gives us column percents, the percents at the bottom are always 100%. Thus, they don’t give us any useful information.

Up in the left-hand corner of the screen you see a menu item, “Graphs”. This lets you display the figures in graphic format. Try it! The “Telegraph Bars” choice gives you the graph on the left; the “Column Distributions” choice gives you the graph on the right.



Neither graph shows any difference between men and women’s percentages. The left-hand graph shows that men and women are almost equal on each row percentage. The right-hand graph shows that their columns are almost exactly alike. This is a further, visual, confirmation that men and women are clearly equally happy.

Close this cross-tabulation, then **type “6”** (Var 6: RACE B/W) in the box for the independent (column) variable and **type “113”** (Var 113: HAPPY) in the box for the dependent (row) variable. Leave the third box blank again, and **click “OK”** to see a table comparing Euro-Americans and African Americans’ happiness. (The table is at the top of the next page.)

Distributions & Cross-Tabulations

The screenshot shows a SPSS Cross-Tabulation window. The independent variable is 'Race B/W (Respondent's race (B/W))' and the dependent variable is 'Happy? (In general - are you very happy - pretty happy - or not too)'. The table displays the following data:

	White	Black	ROW TOTALS
Very	33.5% (740)	24.6% (103)	32.0% (843)
Pretty	57.5% (1273)	58.7% (246)	57.7% (1519)
Not Too	9.0% (199)	16.7% (70)	10.2% (269)
COLUMN TOTALS	100% (2212)	100% (419)	

At the bottom of the window, it states: 'Influence of Var #6 (Race B/W) on Var #113 (Happy?)' with '2531 INTERVIEWEES' and 'Probability > 99.9% (p < .001)'. A 'Close' button is visible in the bottom left corner.

This table compares Whites and Blacks to see if there is any systematic difference in their sense of happiness. And there is! We see that 33.5% of the Whites, but only 24.6% of the Blacks, say they are "very" happy. Only 9.0% of the Whites, but 16.7% of the Blacks say they are "not too" happy. (Remember that you can click and move the left mouse button over any cell to see how to read it.)

Looking at the bottom line, we see that the “Probability > 99.9%” that this difference stems from a real difference in the American population, not from chance. That's quite a difference: there is only 1 chance in a thousand that a difference this big is a result of the particular individuals who ended up in the GSS sample. We can be very sure that fewer Blacks are happy than Whites, and by fairly large margins.

Of course, not every Euro-American is happier than every African American. The GSS figures tell us that a higher percentage of the former than the latter say that they are “very happy”, and a higher percentage of the latter than the former say that they are “not too” happy. Like most sociology, we’re talking about trends here, not individual people.

To Review: In any cross-tabulation, compare the column percents in each row to see whether there are any systematic differences. Use the “Probability” line to tell you how likely it is that any difference you see in the survey sample reflects a real difference in the American population.

DEPENDENT vs. INDEPENDENT VARIABLES

You may be wondering how we know which variable to enter in the rows and which to enter in the columns.

In general, **the row variable is the dependent variable and the column variable is the independent variable**. A variable is independent if it influences or causes another, and is dependent if it is influenced or caused. We first enter the variable that we think will be influenced: the one on which we expect people to differ. We then enter the variable that we think will do the influencing: the one that captures the original differences between people.

In our last example, we wanted to see if the races differed in their reported happiness. Race is the dimension of human difference; we hypothesize that these differences will result in different reports of happiness. And we were proved right: race influences happiness. Race is independent, happiness is dependent, because it would be nonsensical to claim that happiness influences race.

Similarly, in our previous example we treated HAPPY as the dependent variable and SEX as independent. It makes sense for one's gender to determine one's happiness but not the other way around. Unlike race, however, we found that gender makes no difference. Men and women have pretty much the same level of happiness, while Blacks and Whites do not.

A WORD ABOUT "PROBABILITY = N/A"

I think it is pretty clear that sociologists only accept results that are not due to chance – something that we can tell from the "Probability" line that appears at the bottom of every cross-tabulation screen. The probability of the difference between column percents in our sample reflecting a difference in the population as a whole has to be at least 95% for us to claim that those differences are real. This means that 19 of every 20 samples will match the scores for the whole population.

Sometimes, however, you'll get a cross-tabulation that says "Probability = n/a" – i.e.: not available. (You haven't seen any so far, but it's just a matter of time until you do.) What does this mean?

Simply put, "Probability = n/a" results when the sample size is too small for Sociological Insights[®] to figure a chi-square test of statistical significance. If a completely random distribution would put fewer than 5 individuals in each of 20% or more of the cells in the cross-table, then the chi-square can't tell us the likelihood of our particular sample matching the whole population. This doesn't usually happen with large samples – and the GSS is suitably large in most instances. But cross-tabulations between variables that have lots of categories run the risk of having a few cells without many occupants.

Having a "Probability = n/a" does not mean that a cross-tabulation does not reveal any patterns. The pattern may be quite clear. It just means that we don't have a large enough sample for Sociological Insights[®] to tell us whether the column differences are significant or not. We cannot rely on math but have to use our judgment. That is to say, we must be wise in our interpretations. In some cases, the pattern is clear. In others, we can still describe the apparent patterns, while noting that we would need to see a larger sample before claiming that the patterns are characteristic of the population at large.

THINGS TO REMEMBER

1. **Survey data** come from interviews with individuals. Surveys researchers interview a **random sample** from a target population, so that they can generalize the information they get from the sample to the population as a whole.
2. **Distributions** show how many people gave each possible answer to a question.
3. **Cross-tabulations** compare the answers given by different types of people. Enter the types you want to compare as the **column** (independent) variable and the question that you want to compare on as the **row** (dependent) variable.
4. Compare **column percents** to see how the various categories of the independent (column) variable differ from one another. Comparing columns allows one to test hypotheses about the influence of the column variable on the row variable. If the column variable has no influence, then there will be no differences between the columns. If it has an influence, then there will be differences.
5. The **“Probability” must be at least 95%** (and $p \leq 0.05$) **for us to conclude that the difference between the columns reflects a real difference in the population**. If it is less than 95% ($p > 0.05$), then the results we see are due to **sampling error**, and there is **no** actual difference between the columns. In this latter case, the apparent differences are due to the ways that this particular sample differs from the population at large.
6. **“Probability = n/a”** indicates that our sample is not large enough for the program to figure a chi-square test of significance. This occurs when a completely random distribution of people would put fewer than 5 individuals in each of at least 20% of the cells. This does not mean that the cross-tabulations do not reveal patterns. It just means that we must be wise in our interpretations.
7. Cross-tabulations do not let you conclude anything about individuals. They do, however, allow you to describe patterns and tendencies among populations as a whole. The fact, for example, that a higher percentage of Whites than Blacks say that they are happy does not allow you to conclude that every Euro-American is happier than every African American.

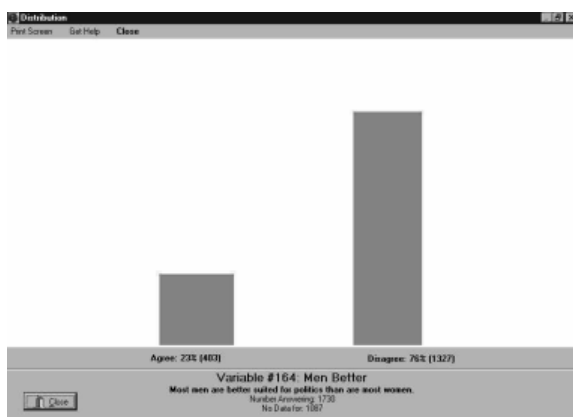
SEVEN: INDEXES

In the 2000 General Social Survey, respondents were asked several questions about gender roles. Among them were:

"Do you agree or disagree with this statement? Most men are better suited emotionally for politics than are most women."

"Do you agree or disagree with this statement? It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family."

Both of these questions measure sexism. They identify those people who think men and women should stick to their traditional roles: man as provider and leader, woman as homemaker. As you might expect, some people held patriarchal attitudes about the sexes, while others were more liberal. But what kind of people held which attitudes?



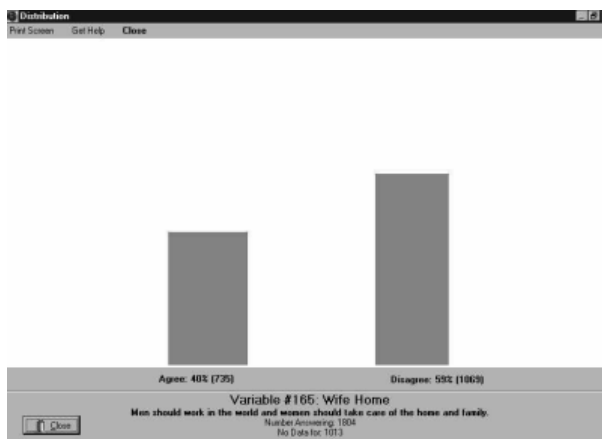
Start Sociological Insights®. Click open the “Survey” menu, then choose “Distribute”. Type “164” (Var 164: MEN BETTER) in the blank and click “OK”. You’ll see the distribution in the chart on the left.

As you can see, 23% of Americans agree that men are better suited to politics, while 76% oppose it. Not so bad, if you believe in gender equality.

Hit <ESC>, then type “165” (Var 165: WIFE HOME) in the blank and click “OK”. You’ll see the distribution below:

This news is not so good. 40% of American adults think that men should be the primary breadwinners and women should have primary responsibility for home and family. This does not necessarily mean that they oppose women working. They may think that working women should also take charge of the childcare, cooking, and housework – a “second shift” that working husbands can neglect.

(The GSS didn’t ask who should do the yard work and fix cars.)



A SEXISM INDEX

This introduces an interesting dilemma. Which one of these questions do we use, if we want to find out what kind of people are sexist in America today? If we use the political question to compare the attitudes of housewives, working wives, and working husbands (Var 29: HSWF/ WKWF), for example, we see that these groups are all very much alike. (“Probability = 50-74%.) If we use the gender role question, however, we see that they are quite different. Housewives

support strict role divisions, working wives do not, and working husbands are somewhere in between. If we want a common measure, which one should we use?

The problem comes, of course, because the 23% who think that women are ill-suited for politics are not all included in the 40% who think that women should take primary charge of the home. There are some people who think that women are perfectly well-suited for politics, but should manage the home, and other people who think that women are ill-suited for politics, but that work and home should not be gender-divided domains. We can see this with a simple cross-tabulation.

Close the distributions chart and hit <ESC> to get back to the main program screen. Open the “SURvey” menu, choose “Cross-Tabulate”, and type “165” (WIFE HOME) and “164” (MEN BETTER) and in the first two boxes. (It doesn’t matter which one goes where.) Leave the third box empty. Now click “OK” to get a table that shows how many people fall into each of the four possible categories:

Independent Variable: Wife Home (Men should work in the world and women should take care of the home and family)							
	Agree	Disagree					ROW TOTALS
Agree	39.0% (260)	13.4% (136)					23.5% (396)
Disagree	61.0% (406)	86.6% (880)					76.5% (1286)
COLUMN TOTALS	100% (666)	100% (1016)					

Dependent Var.: Men Better (Most men are better suited for politics than are most women.)

Influence of Var #165 (Wife Home) on Var #164 (Men Better)
1682 interviewees
Probability > 99.9% (p < .001)

We are not comparing columns, so the percentages don’t mean much here. Ignore them and just look at the raw numbers under each. Out of the 1682 interviewees who were asked this question, 260 agreed with both statements, while 880 disagreed with both. So far, so good: we would certainly consider the former to be sexist and the latter to be non-sexist. But what about everyone else?

542 of our respondents split their votes. 136 of them agree that men are emotionally better suited for politics than are women, but disagree that men should manage the work world and women should manage the home. 406 disagree with the political statement but agree with the gender role statement.

In situations like this, sociologists construct an **index**. An index combines the answers that people give to several related questions. It reveals an underlying pattern that each of these questions alone may obscure.

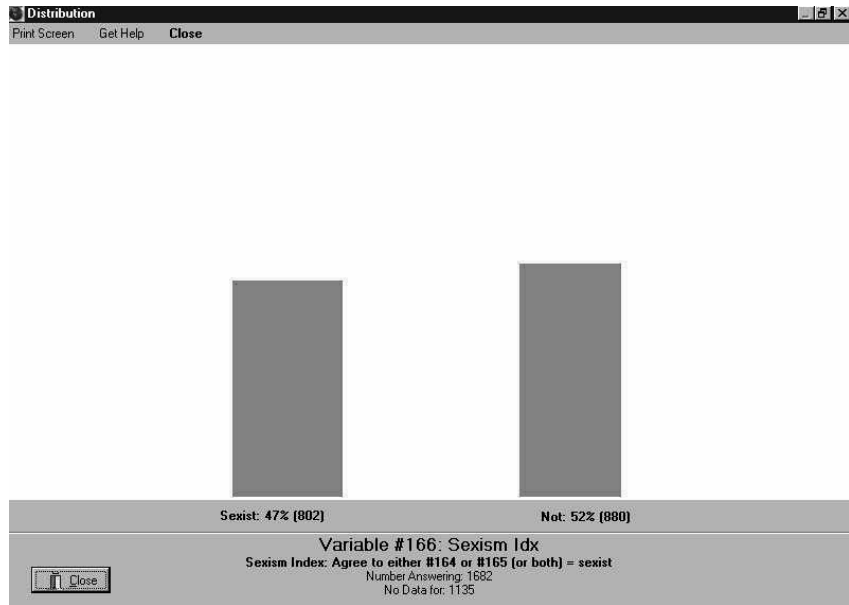
In this case, we can divide people into four different groups:

<u>agree w/ both</u>	<u>agree w/var 164 only</u>	<u>agree w/ var 165 only</u>	<u>disagree w/both</u>
260 (15%)	136 (8%)	406 (24%)	880 (52%)
← more sexist less sexist →			

We can locate these four groups along a sexism scale that runs from 4 (on the left) to 1 (on the right). Clearly, those who agree with both statements get a “four”. Assuming that we think that agreeing with the political question is more sexist than agreeing with the gender role question,* we give those who agree only with the former a “three” and those who agree only with the latter a “two”. Those who disagree with both statements get a “one”. We now have a single index for measuring people’s gender attitudes.

An **index** combines people’s answers to several related questions into a single measure.

The 2000 GSS data set that comes with Sociological Insights[®] goes one step further. Partly in order to reduce the number of categories, and partly to avoid the choice of whether political attitudes or gender role attitudes are “more” sexist, I have collapsed these four categories into two. **Variable 166: SEXISM IDX** assigns any person who agrees with either one of the aforementioned questions to the sexist camp. Non-sexists are those who disagree with both. Here is the distribution:



By this measure, the American population is nearly evenly split. 52% of GSS respondents disagreed with both statements; 47% agreed with one or both of them. To someone who has lived in many parts of the country and among many different kinds of people, this seems roughly right. It does not mean that everyone labeled “sexist” is necessarily misogynist; it simply means that those labeled “non-sexist” have a strong ideal of gender equality in political, work, and home spheres.

* Agreeing that men are emotionally better suited to politics than are women is to say that women lack something that men have. Agreeing that men should be breadwinners and women should manage the home does not imply that either one has greater capabilities. The first of these seems more sexist to me.

WHO IS SEXIST?

Now we can get down to cases. From the opening *Sociological Insights*[®] screen, **click the “Survey” menu, choose “Cross-Tabulate”, and type “166”** (Var 166: SEXISM IDX) in the rightmost (second) box. This will be the dependent (row) variable for the next several tables. We will enter a series of demographic variables as independent (column) variables – sex, race, and religion among them – so that we can see what kind of people believe in gender equality and what kind of people don’t.

First, let’s look at men and women. **Type “4”** (Var 4: SEX) in the left-hand box and **click “OK”**.

	Male	Female
Sexist	48.8% (345)	46.9% (457)
Not	51.2% (362)	53.1% (518)

Probability = 50-74%

As you can see, slightly more men are sexist than women. But look at the “Probability = 50-75%” at the bottom of the screen. This means that there is only a bit more than an even chance that this small difference reflects a real difference in the American population. Remember that we won’t accept anything less than a 95% chance of being right. We can conclude that men and women have similar gender attitudes.

Close the table. Again use **Var 166: SEXISM IDX** as the dependent (row) variable and use **Var 6: Race B/W** as the independent (column) variable. **Click “OK”** to get the table.

	White	Black
Sexist	48.1% (636)	44.7% (123)
Not	51.9% (686)	55.3% (152)

Probability = 50-74%

Again, we see a bit of a difference between the columns, but the probability of it reflecting a real difference in the American population is still less than 95%. We conclude that Whites and Blacks have similar gender attitudes.

Now use **Variable 48: POL PARTY 1** as the independent (column) variable. This variable assigns those “leaning Democrat” and “leaning Republican” to those parties, rather than grouping them with the Independents, as Variable 49: POL PARTY 2 does.

	Dem	Ind	Rep
Sexist	43.9% (334)	44.6% (153)	55.1% (305)
Not	56.1% (426)	55.4% (190)	44.9% (249)

Probability > 99.9%

This time, we’ve found a difference. Republicans are much more sexist than Democrats. They are also more sexist than those who identify with neither political party. Notice that “Probability > 99.9%”, which means that there is only one chance in 1000 that our result is peculiar to the particular groups of people who were interviewed for the 2000 GSS. We can conclude that a higher percentage of Republicans than of Democrats and Independents are opposed to women’s equality. (But 44-45% of the latter are still sexist.)*

Here’s another dividing line: between those born in the USA and those born in foreign countries. Use **Variable 38: BORN US?** as the Independent (column) variable. A bit over a third more foreign-born than native-born agreed with one of our two questions. “Probability = 99.9%”, so we can conclude that Foreign-born Americans are more sexist than are native-born Americans.

Born USA?: Were you born in the USA?		
	Yes	No
Sexist	46.3% (707)	61.1% (91)
Not	53.7% (821)	38.9% (58)

Here are a few more. First, **Variable 8: REGION** and **Variable 34: REGION @16**. Both are significant (Probability > 99.5% and > 99.9% respectively). For both, significantly fewer Southerners than those from other regions believe that women are (or should be) equal to men.

Region : Region where respondent lives				
	NthEast	South	Midwest	West
Sexist	45.0% (157)	53.6% (316)	42.5% (169)	46.4% (160)
Not	55.0% (192)	46.4% (274)	57.5% (229)	53.6% (185)

Region @16: Region lived in at age 16				
	NthEast	South	Midwest	West
Sexist	42.7% (156)	53.9% (205)	44.4% (195)	39.8% (100)
Not	57.3% (209)	46.1% (244)	55.6% (244)	60.2% (151)

* We would have found the same difference had we used Variable 49. It doesn’t seem to matter where we assign those merely “leaning” toward one or another of the major parties.)

Second, let's look at religion. **Variable 96: RELIGION** divides people into Protestants, Catholics, Jews, Other, and None. **Variable 97: FUND/LIB** identifies the relative fundamentalism, moderation, or liberalism of the respondent's chosen religion. **Variable 99: ATTEND CH** divides people by how often they attend church. Which of these groups are sexist?

Religion: What is your religious preference?					
	Prot	Catholic	Jewish	Other	None
Sexist	52.7% (485)	50.6% (201)	27.8% (10)	39.1% (34)	28.9% (69)
Not	47.3% (435)	49.4% (196)	72.2% (26)	60.9% (53)	71.1% (170)

Fund/Lib: Fundamentalist/Liberal denomination				
	Fund	Moderate	Liberal	
Sexist	58.4% (286)	49.0% (289)	37.5% (184)	
Not	41.6% (204)	51.0% (301)	62.5% (307)	

Attend Ch: How often do you attend religious services?					
	Never	LT 1/mo	LT 1/wk	1+/week	
Sexist	45.6% (157)	42.0% (229)	44.8% (117)	56.9% (281)	
Not	54.4% (187)	58.0% (316)	55.2% (144)	43.1% (213)	

All three of these tables are significant; for each, the “Probability > 99.9%” means that the differences we see between the columns in our sample reflect real differences in the whole American population. The tables show us that Protestants and Catholics are equally sexist, but that Jews and those identifying with no religion are much more supportive of women's equality. Religious conservatives are much more sexist than are religious liberals. And those who attend church at least weekly are much more sexist than is everyone else.

(Several of these categories probably overlap. It is not unreasonable to expect that fundamentalist Protestants who attend church frequently are more sexist than liberal Catholics who only attend church once a month. But we will have to wait until the next chapter to see how we can subdivide populations using two variables at once. For now, we can only speculate about such things.)

USING AN INDEX AS AN INDEPENDENT VARIABLE

So far, we have used the sexism index as a dependent (row) variable. That is, we have asked “How do different kinds of people differ in their gender attitudes?” – and have found out by comparing one column to another. More Protestants are sexist than Jews. Equal percentages of men and women are sexist. More Republicans than Democrats and Independents are sexist. Of course, any individual Protestant or Republican may be non-sexist. But these are the underlying social patterns.

Sometimes, however, we want to use an index as an independent (column) variable. We do this when we think that whatever the index measures will influence other attitudes. It makes no sense to ask whether one's beliefs about women's social roles influence one's race. So we don't use the sexism index as the independent variable when we cross-tabulate it with race – or with any

other demographic measure. But it does make sense to ask whether one's gender attitudes make a difference to how people think about divorce.

Cross-tabulate Variable 176: DIV EASY? (“Should divorce in this country be easier or more difficult to obtain than it is now?”) with **Variable 166: SEXISM IDX**. Use the latter as the independent (column) variable and the former as the dependent (row) variable. Here is the result:

The screenshot shows a SPSS Cross-Tabulation window. The title bar reads "Cross-Tabulation". The main title is "Independent Variable: Sexism Idx (Sexism Index: Agree to either #164 or #165 (or both) = sexist)". The dependent variable is "Div Easy? (Should divorce be easier to get - stop as is - or become harder to get?)". The table shows the following data:

	Sexist	Not					ROW TOTALS
Easier	21.6% (162)	28.6% (239)					25.3% (401)
As is	17.4% (131)	26.6% (222)					22.2% (353)
Harder	61.0% (468)	44.9% (376)					52.5% (833)
COLUMN TOTALS	100% (751)	100% (836)					

At the bottom of the window, it states: "Influence of Var #166 (Sexism Idx) on Var #176 (Div Easy?)", "1507 observations", and "Probability > 99.9% (p < .001)".

Clearly, there is a real difference between these two columns. The “Probability > 99.9%” tells us that. A lower percentage of sexists than non-sexists think that divorce should be made easier to obtain than it is now. A much higher proportion of sexists than non-sexists think that it should be made harder. (Note, however, that a majority of the entire population thinks that divorce should be harder to get. Look at the “row totals” column – on the right – to see how the entire population divides.)

Why might this be? One can imagine all sorts of reasons, including the near-paranoid theory that sexist men just don't want women to escape their clutches. A more plausible hypothesis is that what we are calling “sexism” really amounts to a belief in traditional gender roles. Those who favor those roles are more likely to oppose anything that threatens them – and easy divorces certainly do that.

Imagine, for a moment, that you are a housewife, taking care of your home and family while your husband is the family breadwinner. Imagine that you like this role division. Are you going to want to make it easier for your husband to walk out on you? Or are you going to want divorce to be difficult – a last resort for marriages gone terribly wrong? I suspect that you will favor the latter. Every marriage has its rocky moments, and strict divorce laws make it easier for couples to last through them. People who depend on a male/female division of labor are particularly likely to favor laws that protect that division. It's a sensible thing to do.

Let's check this, by cross-tabulating **Variable 29: HSWF/WKWF** (Housewives vs Working Wives) against **Variable 166: SEXISM IDX**. This time we'll look at the table both ways: first with “29” as the independent (column) variable and “166” as the dependent (row) variable, then reversed. As we'll see, these two views tell us slightly different things.

Here is the table seen normally, with SEXISM IDX as the dependent variable. This tells us whether there are any differences in gender attitudes between three social categories: Housewives, Working Wives, and Working Husbands. We see that there are, and that they are significant; the “Probability > 99.9%” tells us that there is only 1 chance in 1000 that the differences between these columns do not reflect real differences in the American population as a whole.

	HseWife	Wk Wife	Wk Husb
Sexist	64.3% (72)	40.5% (113)	46.9% (138)
Not	35.7% (40)	59.5% (166)	53.1% (156)

Specifically, we see that housewives are much more sexist than are the other two social categories. Working wives are a bit less sexist than working husbands, but the big difference is between the two classes of women. One’s role in life certainly influences ones gender attitudes.

But wait a minute! What is to say that role influences gender attitudes rather than the other way around? If a woman believes that women should manage the home, is she not more likely to find a husband and take the role that she thinks is proper? Maybe we need to turn the table around, to see whether there are any differences in the social roles taken up by sexists and non-sexists.

	Sexist	Not
HseWife	22.3% (72)	11.0% (40)
Wk Wife	35.0% (113)	45.9% (166)
Wk Husb	42.7% (138)	43.1% (156)

Redo the cross-tabulation, this time with **Variable 166: SEXISM IDX** as the independent (column) variable and **Variable 29: HSWF/WKWF** as the dependent (row) variable. You’ll get the table to the left.

This view – which also has “Probability > 99.9%” – gives us some new information. First, we see that working husbands make up the same proportion of both sexists and non-sexists. This social role seems unrelated to sexism. Second, we see that more wives work outside the home than work only inside it – also among both sexists and non-sexists. This confirms

the view that the stay-at-home wife is no longer the norm in American society. Third – and most importantly – we see that a significantly larger proportion of non-sexist women than of sexist women work outside the home. Women who hold traditional beliefs about gender roles do work outside the home, but they do not do so as frequently as do women who believe in gender equality. Perhaps they do not do so willingly, but only because they need the money; unfortunately, the 2000 GSS does not let us test this hypothesis. But we have found a real difference between women here. Together, these two tables tell us quite a bit about the intersection of women’s social roles and women’s gender attitudes.*

Such instances are not rare. Though you must be careful to make sense of your assignment of variables to rows and columns – as dependent and independent variables, respectively – there are cases in which one can learn a lot by reversing them.

* Actually, this particular hypothesis would be better investigated by using **Variable 165: WIFE HOME**. Comparing the gender role attitudes of housewives and working wives does not require a general sexism index. But the results are very nearly the same, as a glance at the cross-tabulations of Var 165 and **Variable 29: HSWF/WKWF** shows.

Here's a guide:

The **independent** (column) variable is the one containing the types of people that you wish to compare. The **dependent** (row) variable is the quality on which you want to compare them. Cross-tabulations let you ask "How do the different kinds of people – in each column – differ from each other?" Just compare the column percents – and make sure that "Probability" is at least 95%.

OTHER INDEXES

The Sociological Insights[®] 2000 GSS data set contains several other indexes. Among them are:

variable #	variable name	description
87	SPCH IDX	Index of willingness to restrict free speech
88	TEACH IDX	Index of willingness to restrict university teachers
89	BOOK IDX	Index of willingness to ban books
90	CIV LIB IDX	Index of civil liberties attitudes constructed from indexes 87, 88, 89
108	RACISM IDX	Index of racist attitudes
120	TRUST IDX	Index of how much respondent thinks others can be trusted
134	CONF IDX5	Index of confidence in major institutions (5 categories)
135	CONF IDX3	Index of confidence in major institutions (3 categories)
136	CI W/O GOVT	Index of confidence in non-governmental major institutions
137	CI W/ GOVT	Index of confidence in governmental institutions
236	GREEN IDX	Index of willingness to make personal lifestyle sacrifices for the environment

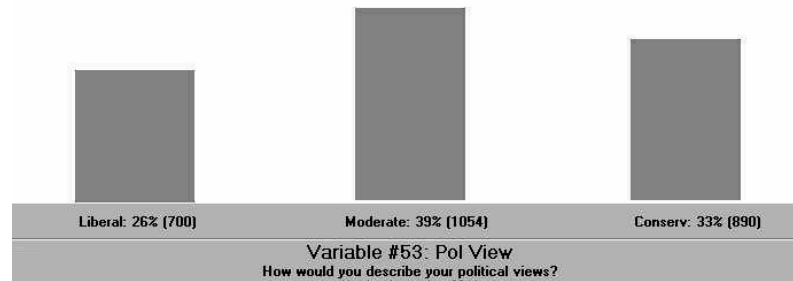
You can work with these indexes just as you worked with the sexism index (Variable 166). Use the index as the dependent (row) variable if you want to see how different kinds of people differ on it. Use it as the independent (column) variable if you want to see how sexism influences other attitudes.

THINGS TO REMEMBER

1. An **index** is a variable that combines scores from several other variables in order to measure a trait that no one of them can measure alone. You can treat it just like any other variable.
2. Usually, you treat an index as a **dependent** (row) variable, to see how different populations differ on whatever it measures. Sometimes, however, you can use it as an **independent** (column) variable, to see how it influences other social attitudes.
3. It is always useful to remember how an index is constructed. Variable 166: SEXISM IDX, for example, labels “Sexist” those who agree with either one (or both) of two specific questions about gender differences. It does not measure misogyny, which is an antipathy to women. It thus helps answer certain social questions but not others.

EIGHT: CONTROL VARIABLES

The General Social Survey asks people about their political views. If you check the distribution of **Variable 53: POL VIEW**, (SUrvey menu, Distribute) you find that 26% of respondents describe themselves as political liberals, 39% describe themselves as moderates, and 33% describe themselves as conservatives.



The cross tabulation of this variable (as the dependent variable) with **Variable 4: SEX** shows that roughly equal percentages of men and women describe themselves as liberals. But more men than

	Male	Female
Liberal	25.0% (292)	27.7% (408)
Moderate	37.0% (433)	42.1% (621)
Conserv	38.0% (445)	30.2% (445)

women say that they are conservatives, and more women than men say that they are moderates. The differences in our sample are also present in the whole population. We know this because the “Probability = 99.9%”.

A similar pattern divides Whites from Blacks. Cross-tabulating **Var 53: POL VIEW** with **Var 6: RACE B/W** gives us the table on the right.

	White	Black
Liberal	26.2% (554)	27.3% (108)
Moderate	38.3% (810)	46.3% (183)
Conserv	35.6% (753)	26.3% (104)

Again, we have roughly equal percentages of Euro-Americans and African Americans describe themselves as politically liberal. More Whites say that they are conservative, while more Blacks say that they are politically moderate. “Probability = 99.9%” tells us that these differences occur in the whole American population.

This leads to an interesting question. Are the gender differences that we find in the whole population equally true of both races? Or do other patterns emerge? Does the numerical dominance of Whites perhaps hide some gender differences among Blacks, that we could see more clearly by looking at Blacks alone?

One can easily answer such questions by using a **control variable**. Basically, a control variable lets us split a population into parts. We then run the same cross-tabulation on each part. This lets us see if social patterns are present in part of a population, whether or not they are present in the whole. We can tell if a pattern is universal, or if it is only true for a subgroup. For the example before us: Are the political differences between men and women true for everyone? Or are they only true for Whites – who make up that population’s largest part? Let’s find out.

Start Sociological Insights[®]. Open the SURvey menu, then choose “Cross-Tabulate”. This time we’ll fill in all three boxes: **type “4” (SEX)** for the independent (column) variable, **“53” (POL VIEW)** for the dependent (row) variable, and **“6” (RACE: B/W)** for the control variable, using **<TAB>** to navigate between them. **Click on “OK”** to run the routine.

The screenshot shows a window titled "Cross-Tabulation w/ 1 Control Variable". The control variable is "Race B/W (Respondent's race (B/W))" with the selected category being "** White **". The independent variable is "Sex (Respondent's sex)" with columns for Male and Female. The dependent variable is "Pol View (How would you describe your political views?)" with rows for Liberal, Moderate, and Conserv. The table shows percentages and counts for each cell, along with row and column totals. At the bottom, it states: "Influence of Var #4 (Sex) on Var #53 (Pol View), controlled by Var #6 (Race B/W). 2117 interviewees answered 'White' on Race B/W. Probability > 99.9% (p < .001)".

	Male	Female	ROW TOTALS
Liberal	29.8% (252)	28.1% (322)	26.2% (564)
Moderate	35.1% (342)	40.9% (468)	38.3% (810)
Conserv	41.0% (399)	30.9% (354)	35.6% (753)
COLUMN TOTALS	100% (978)	100% (1144)	

Let’s take a tour of this screen. It’s pretty much like the cross-tabulation screens that we’ve seen so far – except that it only counts the White part of the population. That’s 2117 individuals, as we see from the line near the bottom. (The “White” button near the top is in larger letters and has “**” around it. This tells us that the cross-tabulation is for the White population only.)

Like the other cross-tabulations we’ve seen, the independent variable (SEX) is at the top and the dependent variable (POL VIEW) is on the left; the marginals are at the right and at the bottom. Reading down the columns, the table shows us how White men’s political views differ from those of White women. And they do: there is only 1 chance in 1000 that the difference between these columns results just from the luck of sampling. We can thus conclude that White men and White women have different patterns of political views – just as do men and women in general.

A **control variable** lets us split a population into parts. By running the same cross-tabulation on each part, we can see whether social patterns are present in the separate parts of a population. We can thus tell if a pattern is universal, or if it is only true for one or another subgroup.

What are these patterns? With 99.9% assurance, we can say that more White women than White men have liberal and moderate political views, and more White men than White women have conservative political views.

Note that this pattern is not quite the same as it was in the general population. Though the moderate and conservative comparisons are the same as before, we see that a higher percentage of White women than of White men call themselves liberals. That is: we find gender differences in all three political persuasions among Whites, where only two of them differed in the population as a whole. This is an important difference.

Yet, discovering it raises a further question. If more White women than White men are liberals, how is it that the same proportion of each sex in the population as a whole sees itself as liberal? Whites make up the bulk of the U.S. population. Is it conceivable that the percentage of liberals among minorities is so different as to shift the whole? Are minority men so much more likely than minority women to be liberals that it makes up for White women's greater liberality? To find out, we'll have to look at the other main part of the population – African Americans.*

Click on the “Black” button. This shows the cross-tabulation of POL VIEW vs SEX for the 395 African Americans in our sample.

	Independent Variable: Sex (Respondent's sex)		ROW TOTALS
	Male	Female	
Liberal	31.7% (44)	25.0% (64)	27.3% (108)
Moderate	47.5% (66)	45.7% (117)	46.3% (183)
Conserv	20.9% (29)	29.3% (75)	26.3% (104)
COLUMN TOTALS	100% (139)	100% (256)	

Influence of Var #4 (Sex) on Var #53 (Pol View), controlled by Var #6 (Race B/W)
395 interviewees answered "Black" on Race B/W
Probability = 75-89% (p = .25-11)

This table is like the last one, except that the text on the “White” button has shrunk and its asterisks have disappeared. The text on the “Black” button has grown and it is now surrounded with “**”. This tells us that we are viewing a cross-tabulation for the African American population only.

What is happening here? Reading just the percentages, it does indeed seem that a higher percentage of Black men than of Black women are politically liberal. It also seems that a higher percentage of Black women than of Black men are politically conservative. The differences between the columns are rather startling. This seems to answer our question in the affirmative: Black men's high levels of liberality seem to balance those of White women – such that we see no difference between men and women when we combine the races.

Maybe – but we must be careful. Look at the “Probability”. It is only “75-89%”. This tells us that there is no real difference between men's and women's political views in the African American population. Though it seems on the surface that African American men are 25% more likely to be liberal than are African American women, and that AA women are more than 40% more likely to be conservative than AA men, this is an artifact of the particular people who were interviewed. The difference is only apparent, not real!

* Whites and Blacks are such a large part of the GSS sample that those listed as “Other” make no difference on this score. I have thus left them out. This makes explaining easier. Feel free to check this by controlling your cross-tabulation with **Variable 5: RACE** in place of **Variable 6: RACE B/W**.

The problem is that the African American sample is too small. Remember that the likelihood that particular differences in column percents reflect a real difference in the population depends on the sample size. Were we to get these same percentages from a sample of 2000 people, we would know that Black men and women are highly different in their political views. But our sample contains only 395 people – the number of African Americans in the GSS pool who gave us both their political views and their gender. This is not enough to ensure that the sample represents the whole Black population. More exactly, we cannot be at least 95% sure that it represents the whole population; the “Probability” tells us that we are only 75-89% sure. This is not good enough for sociologists, so we reject the hypothesis of a difference between the political views of Black men and Black women. We conclude that Black men and Black women have about the same distribution of political views.

Where does this leave us? Let’s go back to our starting question: Are the political differences between men and women that we saw in the whole population true of both races? These were, for the whole population:

1. the same percentage of each sex are liberals
2. a higher percentage of men than women are conservative
3. a higher percentage of women than men are moderate

Clearly, this is not the case either for Whites or for Blacks. Among Whites, more women than men are liberal, though points 2 and 3 are still true. Among Blacks, only point 1 is true; points 2 and 3 are false – because there is no significant difference between Black men and Black women. How, then, do these figures combine to produce the pattern in the population as a whole?

Let’s take each of our three points, starting with #2 and #3. Given Whites’ numerical dominance, both in the GSS sample and in the population, it makes sense that their pattern would rule, despite the fact that Black men and Black women are roughly equal in their percentages of political moderates and political conservatives. The gender difference among Whites is so large that the gender equality of Blacks cannot change things. Thus, White women’s tendency toward political moderation and White men’s tendency toward political conservatism is reflected in the population as a whole.

However, in the case of political liberalism – our point #1 – a strange thing is happening. Yes, a significantly higher percentage of White women than White men are liberal. But the tendency of Black men toward greater political liberality than Black women tips the entire population toward gender equality – even though the gender difference among Blacks does not rise to the level of statistical significance. That is: our sample contains just enough more liberal Black men than liberal Black women to raise the total number of liberal males to the point that women’s greater liberality vanishes. A percentage difference that is not statistically significant among a small group erases a statistically significant percentage difference in a larger group when the two groups are combined!

This seems strange, yet it is a direct result of the mathematics of probability sampling. Just be thankful that we don’t have to do any math in this book!

We can thus conclude that:

- A higher percentage of White women than White men consider themselves to be politically liberal
- There is no significant difference between the percentage of men and women who consider themselves to be politically liberal – either among Blacks or among the total population.
- White women – but not Black women – are more likely than the men of their race to count themselves as political moderates.

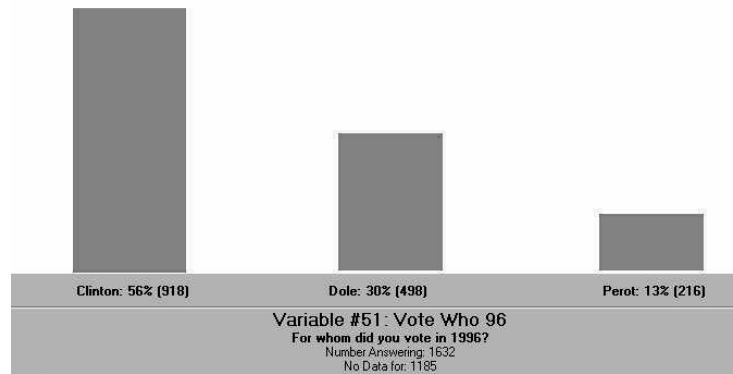
- White men – but not Black men – are more likely than the women of their race to count themselves as political conservatives.

Clear as mud? Think it through slowly and carefully. Such apparently contradictory patterns are common in survey research. One must take special care to tease out the exact relationships that the data show us. They also point us in the direction of further research. It would be helpful to ask these questions of a much larger sample of African Americans, to see whether the pattern of greater male liberality holds true.

In any event, control variables are a powerful tool for discovering patterns hidden in our data. They greatly expand our understanding of what is going on.

A SECOND POLITICAL EXAMPLE

Let’s look at a second example: voting in 1996. The 2000 General Social Survey asked voters whether they voted for Bill Clinton, Bob Dole, or Ross Perot. (They couldn’t ask about voting for Al Gore or George Bush, as the 2000 election had not been held yet.) Here’s the distribution:



Of the 1632 interviewees who were asked this question, 56% said that they voted for Clinton, 30% said that they voted for Dole, and 13% said that they voted for Perot. (They may not have been telling the truth, as some people like to claim to have always backed the winner. But these were their answers.)

Men and women voted differently, however, as did Whites and Blacks. To see how, create two cross-tabulations, using **Variable 51: VOTE WHO 96** as the dependent (row) variable for both. First use **Variable 4: SEX** as the independent (column) variable; then use **Variable 6: RACE B/W**. You will get these two tables:

	Male	Female
Clinton	47.8% (344)	62.9% (574)
Dole	37.7% (271)	24.9% (227)
Perot	14.5% (104)	12.3% (112)

	White	Black
Clinton	49.1% (664)	94.1% (224)
Dole	35.4% (479)	3.8% (9)
Perot	15.5% (209)	2.1% (5)

Both tables show significant differences between the columns: “Probability > 99.9%”. As you can see, a larger percentage of women and African Americans voted for Clinton – overwhelmingly so among the latter. These are the famous gaps that were talked about so much in the late 1990s. The

Control Variables

“gender gap” describes a political divide between men and women. The “race gap” described a similar divide between Blacks and Whites. The race gap was much larger, which is what made the denial of voting rights to Blacks in the 2000 Florida presidential election so damaging to Democrats. But the gender gap was more important nationally, because women are a much larger part of the population.

How do these figures interact? What happens if we compare men’s and women’s voting patterns separately for the two races? What happens if we compare Blacks’ and Whites’ voting patterns separately for men and for women? Let’s find out.

Return to the cross-tabulation variable entry screen, and **type “4” (SEX)** for the independent (column) variable, **“51” (VOTE WHO 96)** for the dependent (row) variable, and **“6” (RACE: B/W)** for the control variable, using **<TAB>** to navigate between them. **Click on “OK”** to run the routine. Push the “White” and “Black” buttons to see both tables.

WHITES

	Male	Female
Clinton	41.9% (259)	55.2% (405)
Dole	42.1% (260)	29.8% (219)
Perot	16.0% (99)	15.0% (110)

Probability = 99.9%

BLACKS

	Male	Female
Clinton	92.8% (77)	94.8% (147)
Dole	3.6% (3)	3.9% (6)
Perot	3.6% (3)	1.3% (2)

Probability = n/a

Clearly, a higher percentage of White women than White men supported Bill Clinton; in fact, if only White men had voted, Bob Dole would have become the President. This demonstrates what we saw before: the larger percentage of conservatives among White men relative to White women. Among Blacks, however, “Probability = n/a”. This is because the sample is too small for the program to figure a chi-square. A purely random distribution – one that has no relationship between the variables – must put at least 5 people in at least 80% of the cells in order for the math to work. Too few of our sample’s African Americans voted for Perot to reach this threshold.

Can we say anything about the relative voting preferences of Black men and Black women? In this case, we can. Comparing the columns, we find almost the same percentages at each point. Among the 238 African Americans who answered this question, nearly identical percentages of men and women voted for Clinton, Dole, and Perot. The numbers are so close, in fact, that we can be reasonably sure that a larger sample size would produce a “Probability ” that confirms their equality. Thus we can conclude that the gender gap is a White phenomenon, not matched in the Black population.

“Probability = n/a” means that our sample is not large enough to figure a chi-square – the statistical test that tells us whether the difference between the columns can be generalized to the population at large. This does not always mean that we can’t generalize this difference. It just means that we have to rely on our judgment, not on math.

(One must be very careful in interpreting cross-tabulations where “Probability = n/a.” Such a result really tells us that our sample is too small to draw firm conclusions. The best solution is usually to increase the sample size, but this is not often possible after survey data have been collected. Another good course is to combine columns or rows, pushing the values above the 5-person threshold. We could, for example, combine the Dole and Perot voters: a “Clinton vs. everyone

else” table would likely give us a useable “Probability” figure. SPSS®, Microcase®, and other professional statistics programs allow you to do this, but Sociological Insights® does not. It does, however, allow two other solutions, which we shall consider in the next two chapters.)

Now turn things around. **Close** this cross-tabulation, then **type “6”** (RACE B/W) for the independent (column) variable, **“51”** (VOTE WHO 96) for the dependent (row) variable, and **“4”** (SEX) for the control variable, using **<TAB>** to navigate between them. **Click on “OK”** to run the routine. Men are on the left and women are on the right.

MEN

	White	Black	
Clinton	41.9% (259)	92.8% (77)	
Dole	42.1% (260)	3.6% (3)	
Perot	16.0% (99)	3.6% (3)	

Probability = 99.9%

WOMEN

	White	Black	
Clinton	55.2% (405)	94.8% (147)	
Dole	29.8% (219)	3.9% (6)	
Perot	15.0% (110)	1.3% (2)	

Probability = 99.9%

A much higher percentage of Black women than White women voted for Bill Clinton. And the same is true for Black and White men. Clearly, the race gap is present for both sexes, even though the gender gap is present only for Whites. Without controlling for race and gender, we cannot see these patterns. By using race and gender as control variables, we can learn much more.

What is the point of controlling cross-tabulations? The technique allows us to examine our data in much greater detail. By splitting a population into parts and running the same cross-tabulation on each, we can see whether social patterns are present in the separate parts of a population. We can thus tell if a pattern is universal, or if it is only true for one or another subgroup.

THINGS TO REMEMBER

1. **Control cross-tabulations** in order to locate differences in the various parts of a population. Are there different relationships between independent and dependent variables for different demographic groups? Use **control variables** to find out.
2. Read controlled cross-tabulations the same way that you read regular cross-tabulations. Use the “Probability = ” figures to tell whether you can generalize the difference between columns from the sample to the population as a whole.
3. Controlled cross-tabulations often produce **“Probability = n/a”**. This happens when a subdivided sample is not large enough to figure a chi-square – the statistical test that tells us whether the difference between the columns can be generalized to the population at large. This does not always mean that we can’t generalize this difference. It just means that we have to rely on our judgment, not on math. (Consult the next two chapters for some other ways to handle such situations.)
4. All statistical analyses require careful thinking, but controlled cross-tabulations require more care than most. Think about what the data tell you before jumping to conclusions.

Control Variables

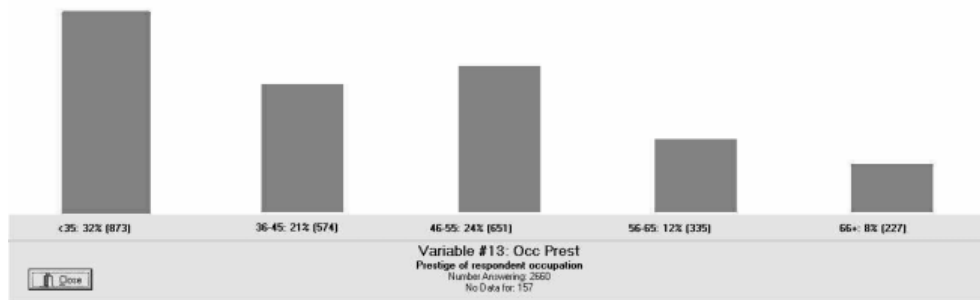
NINE: T-TESTS AND THE ANALYSIS OF VARIANCE

Cross-tabulations are wonderful tools, but they don't always work the way we want them to. Sometimes, there just aren't enough people for *Sociological Insights*[®] to tell us whether two variables have any relationship. This chapter gives you one way around this difficulty. Chapter 10 will give you another. But first we ought to review the problem.

We'll start by looking at occupational prestige. This is a measure of how much respect people give to those in various occupations. In theory, the scale runs from 1 to 100, though in practice it goes from about mid-teens to the mid-80s. Physicians score 86, dentists and college professors 74, accountants score 65, librarians and social workers in the low 50s, bank tellers and nurse's aides in the low 40s, truck drivers 30, garbage collectors 28, and so on.

Sociologists didn't just make up these numbers. They result from years of polling data that ask people of all kinds how much respect they grant to various ways of earning a living.* (Lawyers rank near the top, at 75; I guess all those lawyer jokes don't reflect what people really think.)

Start *Sociological Insights*[®]. Click on the Survey menu, then choose "Distribute", and type "13" (Occupational Prestige) to get the following chart:



We see that Americans' occupations are shifted toward the bottom end of the scale. Thirty-two percent fall in the bottom third of the ranking, 8% are in the top third, and more than half of the remainder are below the scale's midpoint. This merely tells us that our society has more people at the bottom of the heap than at the top – no news to sociologists.

As we would expect, people with high levels of education tend to have high-ranked occupations – something that you can see for yourself by cross-tabulating Variable 16 (DEGREE) against Variable 13 (Occupational Prestige).

For this exercise, however, we want to look just at the African American portion of the GSS sample. To do this, **choose "Cross-tabulate"** from the SURvey menu. Then **use Variable "16"** (DEGREE) as the independent (column) variable, **use Variable "13"** (Occupational Prestige) for the dependent (row) variable, and now **use Variable "6"** (RACE B/W) as a control variable. **Click the "Black" button** in the upper right-hand corner, under the control variable label. You should get the table at the top of the next page.

* For details, see James A. Davis, *Social Differences in Contemporary America*, San Diego: Harcourt Brace Jovanovich, 1987, pp. 51ff and James A. Davis, et al, *General Social Surveys, 1972-2002: Cumulative Codebook*. Chicago: NORC, 2003.

T-Tests and Analysis of Variance

	Not HS	HS Grad	J Coll	College	Grad Deg
<35	68.9% (62)	46.1% (105)	13.0% (3)	18.4% (7)	0.0% (0)
36-45	20.0% (18)	26.3% (60)	30.4% (7)	13.2% (5)	5.9% (1)
46-55	8.9% (8)	13.2% (30)	17.4% (4)	36.8% (14)	35.3% (6)
56-65	1.1% (1)	11.4% (26)	17.4% (4)	15.8% (6)	35.3% (6)
66+	1.1% (1)	3.1% (7)	21.7% (5)	15.8% (6)	23.5% (4)

This is pretty straightforward: you can clearly see that African Americans who hold a graduate degree are much more likely to have high-prestige occupations than are African Americans who don't have a high school diploma. But take a look at the bottom of the screen:

396 interviewees answered "Black" on Race B/W
Probability = n/a (too many low cells) (p = n/a)

The 2000 GSS only has data on the education and occupational prestige of 396 African-Americans. Only 17 of these have graduate degrees, so the computer cannot use the chi-square test of significance to tell us whether the differences we see between the columns reflect a real difference in the general population. We can guess that this table shows a relationship between education and social class – and it so happens that we would be right to claim one – but we cannot prove the relationship mathematically with this sample. We just don't have enough people.

Let's take another example. **Run the same cross-tabulation**, this time **using Variable 47** (Family Income) as the dependent (row) variable. **Use "16"** again as the independent (column) variable and **use "6"** as the control variable. Again **click on the "Black" button** :

	Not HS	HS Grad	J Coll	College	Grad Deg
< \$15K	68.2% (60)	33.0% (67)	10.0% (2)	8.8% (3)	6.3% (1)
\$15-34K	23.9% (21)	34.0% (69)	50.0% (10)	26.5% (9)	12.5% (2)
\$35-59K	6.8% (6)	23.6% (48)	25.0% (5)	20.6% (7)	12.5% (2)
\$60-109K	0.0% (0)	8.9% (18)	15.0% (3)	38.2% (13)	37.5% (6)
\$110K+	1.1% (1)	0.5% (1)	0.0% (0)	5.9% (2)	31.3% (5)

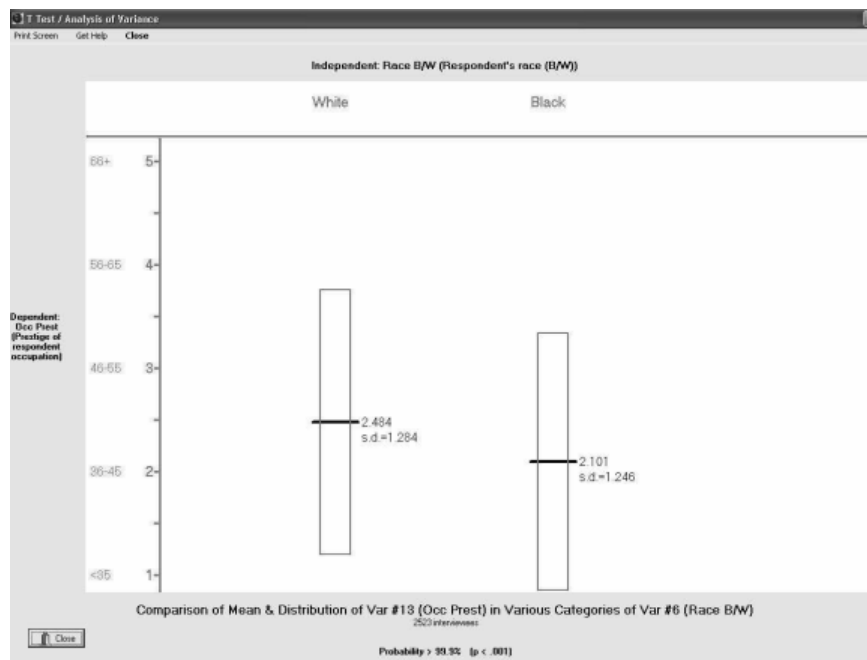
This table also shows a clear positive relationship between education and family income, but again we see that "Probability = n/a". The cells in both the upper-right and lower-left corners of the table

have too few people in them for Sociological Insights[®] to use the chi-square test of significance. We cannot prove the relationship, even though we are pretty sure that one exists.

Fortunately, there is a way around this difficulty. Cross-tabulations and chi-squares are not the only tests that show whether differences between a sample's columns reflect differences in the underlying population. This chapter will explore two other tests for such matters: **t-tests** and **ANOVAs** (which stands for "**analysis of variance**"). Under certain circumstances, these tests can solve our problem.

COMPARING MEANS

Go back to Sociological Insights[®] opening screen, **choose the SUurvey menu**, then **choose the "TEst for Similar Means" menu item**. Type "6" (RACE B/W) in the independent variable (column) box and type "13" (Occupational Prestige) in the dependent variable (row) box. For now, leave the third box empty. Click <OK> to see the following **box plot**:



Let's see how to read this. At the top of the chart, we have our independent variable, with its various categories arranged in columns, as before. On the left, we have a scale that runs from 1 at the bottom to 5 at the top. Each of these numbers has a label that locates the occupational prestige rankings that we saw previously. One stands for those ranking less than 35 (i.e., lower than child care workers); five stands for those ranking 66 and above (high school teachers and nurses). In between, we have (moving up from the bottom) three groups encompassing barbers to bank tellers, insurance agents to librarians, and nursery school teachers to accountants.

Rather than grouping these in table cells, however, Sociological Insights[®] has worked on them mathematically. For each column, it has figured the average (mean) score, which is marked by the black line. It has also figured the standard deviation of each column and drawn a box that ranges from one standard deviation above the mean to one standard deviation below it. Typically, about 68% of the people fall between one standard deviation above the mean and one standard de-

viation below it. This is not always the case, but it happens often enough that **box plots** like these are traditionally drawn this way.*

As you can see, the average White score is 2.484, compared to an average Black score of 2.101. African Americans in our sample, on average, appear to have lower-prestige occupations than do White folks. Sixty-eight percent of them have jobs that range in prestige from waitress to mail carrier; 68% of Whites have jobs that range in prestige from receptionist to airline pilot. (See the note on page 95 if you want to see the whole occupational prestige scale.) The two groups overlap considerably, but the average White still has a higher status job than the average Black.

T-tests compare two groups, to see whether their means differ.

ANOVAs compare more than two groups, to see whether their means and standard deviations differ.

This, however, is just a sample. The key question is whether the two columns are different enough that we can be at least 95% sure that the difference we see in our sample reflects a real difference in the American population as a whole. This amounts to asking whether the means are significantly different from each another, given the size and variability of the sample itself.

Social scientists typically answer such questions using one of two related statistics. When there are only two categories for the independent variable, we use a “**t-test**”. (This test is also called “Student’s *t*”, after the beer brewer who developed it to regulate his brewing. He then published the technique, using the pseudonym “Student”.) When there are more than two columns, we use an “**analysis of variance**” – typically shortened to ANOVA. Both are relatively easy to calculate, based on the means and standard deviations of the scores in each column.**

These tests are handy because both work on considerably smaller samples than are needed for good cross-tabulations. In fact, the “t-test” works anytime that the sample is larger than about 30. This is almost always the case with GSS data! These tests thus avoid the problem that we ran into with cross-tables, which broke up our sample into lots and lots of cells. They test the significance of the sample as a whole, rather than depending on 80% of the table’s cells having at least 5 people in them.

(The only limiting factor is that our dependent variable must be either ordinal data or interval/ratio data. T-tests and ANOVAs don’t work with categorical dependent variables – and they also don’t work with interval/ratio independent variables. We’ll take up this issue later in the chapter.)

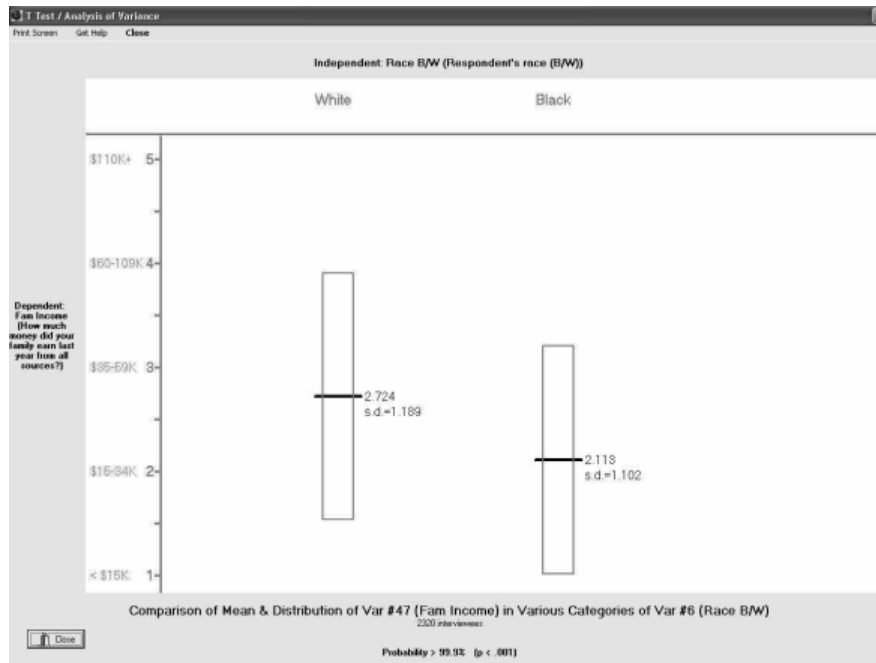
Sociological Insights[®] doesn’t show you the math, but it does show you the results. The bottom of our screen shows the number of people in our sample and calculates the probability that the difference between the columns reflects a real difference in the population as a whole. In this case, with 2523 interviewees, we can be more than 99.9% sure that Whites and Blacks differ from one another on occupational prestige, and furthermore that they do so in the direction that our sample indicates. Whites, on average, have higher occupational prestige than do African Americans.

Now make a similar test of the relationship between race and family income. **Type “6”** (RACE B/W) in the box for the independent (column) variable and **type “47”** (Family Income) in

* A box-and-whiskers diagram adds lines at the top and bottom to show the total range of values. GSS samples, however, almost always cover the entire range, making the whiskers part of the diagram irrelevant. Both box plots and box-and-whiskers diagrams show the mean and standard deviation, however.

** The calculations behind the **t-test** generate a **t-statistic**, which Sociological Insights[®] compares to an established table of values to see whether the columns are significantly different from one another. An ANOVA similarly generates an **F-ratio**, which is compared to a different table. The logic is the same in the two cases, but the calculations are different.

the box for the dependent (row) variable. Leave the third box empty. **Click <OK>** to see the following screen:



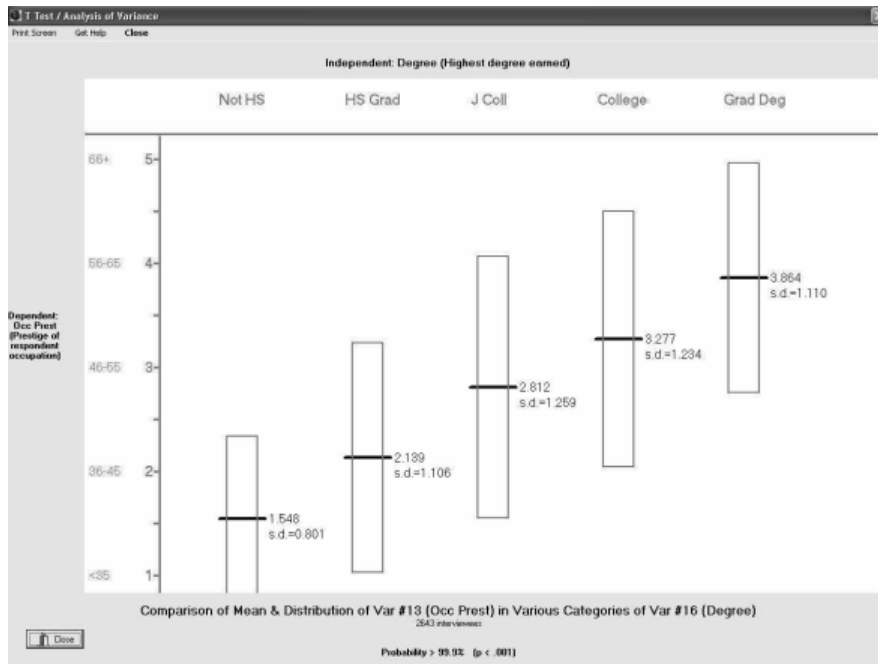
This box plot looks very much like the last one, except the numbers are different. The GSS provides income figures for only 2320 individuals. The White mean is 2.724, which corresponds to about \$45,000. The African American mean is 2.113, which corresponds to about \$30,000. These figures are close to, but do not exactly match, U.S. Census Bureau figures for the same year.*

Again, despite the fact that many African American families in our sample make more money than do many White families, the average White family has more income than does the average Black family. Running a t-test produces the “Probability” calculation at the bottom of the screen, which tells us that we can be 99.9% sure that the difference that we see here reflects a real difference in the American population.

MULTIPLE COLUMNS

Let’s go back to the example with which we began this chapter – of the relationship between education and occupational prestige. **Choose Variable 16 (DEGREE)** as the independent variable, then **choose Variable 13 (Occupational Prestige)** as the dependent variable. Leave the control box empty and **click OK** to get the chart on the top of the next page.

* They actually match the Census Bureau’s median family income figure better than they match the mean income figures for each racial group. This is because a very small number of very wealthy people significantly raise U.S. mean income – and such people are seldom captured in a 2300-person survey. Furthermore, the GSS asks people to locate their income in ranges: \$10,000-\$12,499; \$12,500-\$14,999, ..., \$25,000-\$29,999; etc., all the way up to over \$110,000. This truncates the top of the income scale. Combine this with Americans’ tendency to understate their income, and we are left being better able to compare the relative standing of various groups than to establish any one group’s actual income level. No matter. T-tests and ANOVAs are all about comparison.



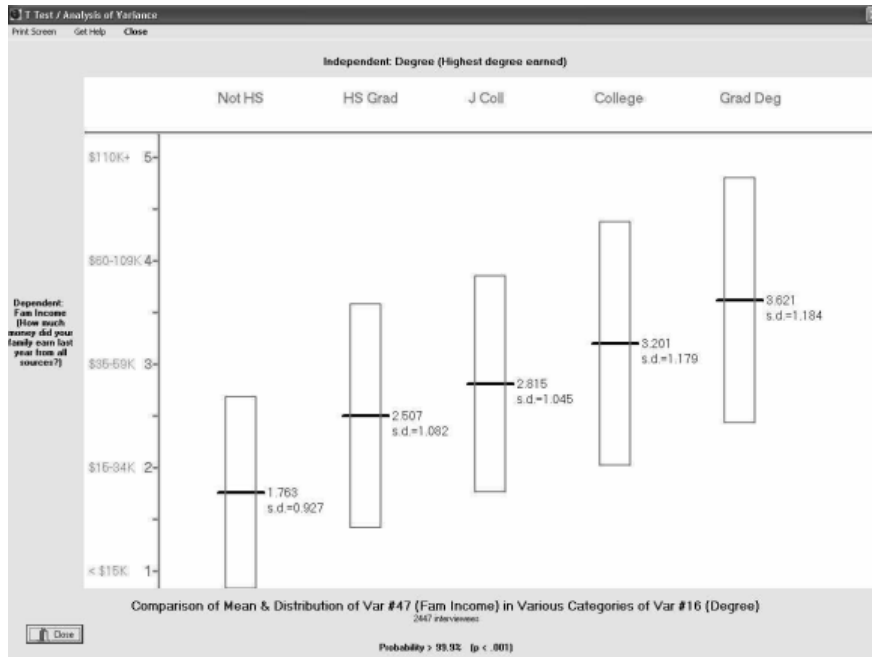
This time, Sociological Insights[®] has run an **analysis of variance** rather than a **t-test**. That’s because a t-test only works with variables that have just two categories and Variable 16: DEGREE has five. From the reader’s point of view, however, there’s little difference. One reads both graphs the same way.

In this case, we can see a clear difference between the column means. As education goes up, so does occupational prestige – and quite dramatically. The mean prestige for those with college and graduate degrees is more than twice that of the mean for those without a high-school diploma. There is, in fact, no overlap between the left-most and right-most columns. Proportionally speaking, very few uneducated people have high-prestige jobs, and very few educated people have low-prestige ones.

Hovering your mouse over the “Probability” line tells you that the columns are significantly different. In fact, there is just one chance in a thousand that our sample of 2643 interviewees does not reflect real differences in the American population. We can be sure that education matters.

“Probability” and “p” mean exactly the same thing for t-tests and ANOVAs as they do for cross-tabulations.

Close this screen, then use **“16”** (DEGREE) as the independent variable, use **“47”** (Family Income) as the dependent variable, and leave the control variable space blank. **Click “OK”** to get the diagram on the top of the next page.

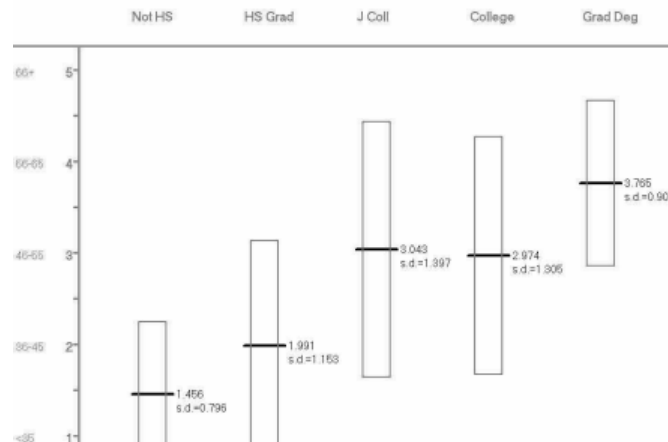


This box plot is very much like the last one. The means and standard deviations are different, of course, but the pattern is the same: more education is systematically connected to higher family income – and there is only 1 chance in 1000 that this is due to the particular people who happened to be included in the 2000 GSS sample. We can be almost sure that education, family income, and occupational prestige all go together.

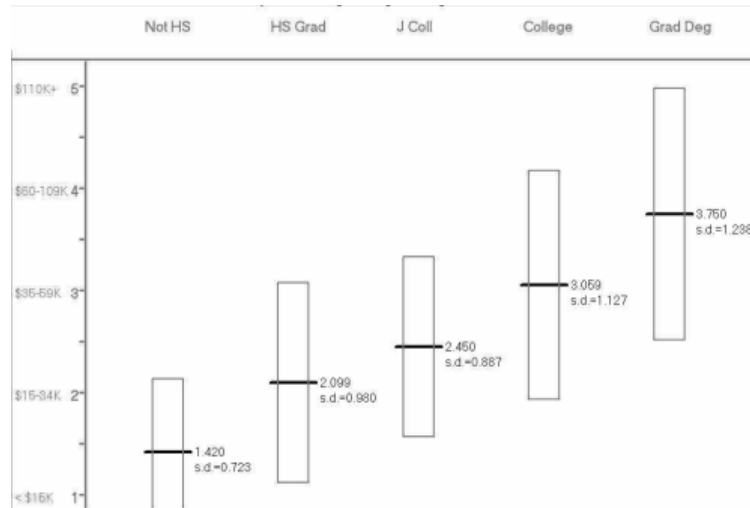
ADDING CONTROL VARIABLES

The question with which we began this chapter, however, was about African Americans, not all Americans. Cross-tabulation was able to demonstrate this relationship among the whole population as well as among White people. It is only among African Americans that cross-tables and the chi-square test of significance failed. So we need to use race as a **control variable** with our ANOVA, to see if it can test the significance of our sample where the cross-tables could not.

Use **Variable 16 (DEGREE)** as the independent variable, **variable 13 (Occupational Prestige)** as the dependent variable, and **variable 6 (RACE B/W)** as the control variable. Click **“OK”**, then **click the “Black” button** to get the following:



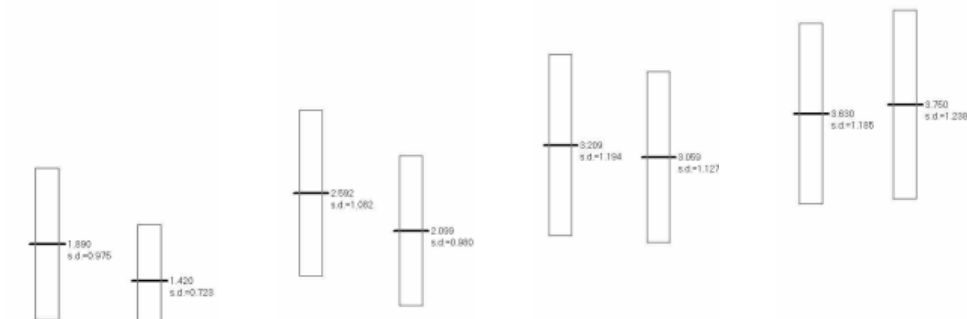
I've truncated the graph to save space, but the result is clear. There is a clear upward movement from left to right, broken only by the relative occupational equality of those who just attended college and those who actually graduated. The "Probability" line shows that we can be 99.9% sure that the pattern that we see in this sample of 396 African Americans is reflected in the whole African American population. If you **close** this screen and **type in "47"** (Family Income) as the dependent variable – making no other changes – we get pretty much the same result for Blacks:



If anything, education is even more influential in producing income than it is in producing occupational prestige. And comparing the charts for the two races seems to indicate that education does a better job of raising Black income levels than it does in raising White income. Of course, Blacks start lower. The four left-most categories show that, on average, college-educated Blacks make less than college-educated Whites, that Blacks with high-school degrees make less than Whites with high-school degrees, and that Blacks who did not graduate from high school make less than Whites who similarly did not graduate. Only at the graduate degree level does the average Black family income seem to exceed that of their White counterparts.

Control variables work exactly the same way with t-tests and ANOVAs as they do with cross-tabulations. Like any independent variable, a control variable must be categorical.

We should test this, of course – and doing so shows us that this impressionistic picture is partly right and partly wrong. **Reverse the independent and control variables by typing "6" in the independent box, "47" in the dependent box, and "16" in the control box.** Suitably truncated, here are the results (left to right) for those without and with a high school diploma, those with a college diploma, and those with a graduate degree. (I've left out those with some college who didn't get a 4-year degree.) In each pair, Whites are on the left and Blacks are on the right.



Viewed side by side, it is clear that there is more difference between Whites and Blacks at the lower educational levels than there is higher up the scale. In fact, only the two left-most pairs are different enough to let us extrapolate from the GSS sample to the population at large. The White-Black income differences at the college and graduate level have a less than 80% certainty of reflecting the wider population. That's not good enough for sociologists! We have to conclude that the disparity in income levels between the races only applies to those with less than a college degree. Education levels make no difference to family income for those who have graduated from college. This means that the White/Black income gap vanishes at higher education levels – a good advertisement for educational attainment!

For both races, in fact, higher levels of education translate into higher family income. Not for everyone, of course, because we are dealing with averages. But that is the picture we see.

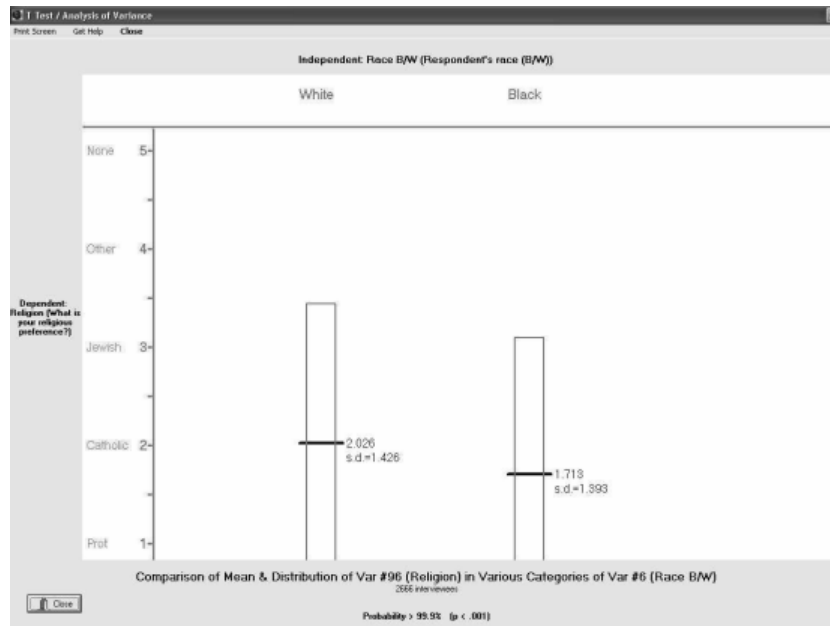
A WORD ABOUT DATA

Clever readers will have noted something about the preceding discussion. In each instance, the independent variable was categorical and the dependent variable was interval/ratio. The reason ought to be obvious: t-tests and ANOVAs require us to create means and standard deviations on the dependent variable. We can absolutely do this with income and occupational prestige, because these are numerically based rankings.* We cannot, however, reasonably do this with such categorical variables as sex or race. What is the American population's average sex? Its average race? Its average marital status? Its average religion? Such questions are meaningless, because sex, race, marital status, and religion are not ranked. We can speak of the **modal** sex, race, etc – i.e., the one that occurs the most often. But one can't do math with modes. T-tests and ANOVAs depend on math to work.

Unfortunately, this is where we run up against one of the basic limitations of computers. Computers are really dumb. A computer stores sex, race, marital status, and religion as a series of ones and zeros, and it is perfectly happy to do math on them. The result is meaningless, but the computer doesn't know that. So we have to make sure that we put the right kinds of data in the right place.

Here's a dreadful example. At the "T-Test and ANOVA" variable entry screen, **use "6"** (RACE B/W) for the independent variable, then **use "96"** (RELIGION) as the dependent variable. Skip the control variable and **click "OK"** to get the box plot at the top of the next page.

* Technically, Sociological Insights[®] has collapsed these to rank-ordered groupings – i.e., to ordinal data. This produces slightly different means and standard deviations than would be the case were we using the raw GSS data. It does not, however, alter either the underlying relationships or their statistical significance.



According to the computer, the “average” White person is Catholic, the “average” Black person is partway between Catholic and Protestant, and the standard deviation of each ranges from fully Protestant to past Jewish, heading toward other. And this difference is 99.9% likely to match the entire U.S. population.

What meaningless gibberish! A cross-tabulation of the same variables (left) shows that Protestants form a majority of the White population and an even larger majority of the Black population. Catholics make up a quarter of the White population but less than 10% of the black population. Other groups are smaller, except for those claiming “no religion”, which make up about 1/8th of the population as a whole. (As a fun exercise, see what percentage of those claiming “no religion” – Variable 96 – never pray – Variable 102.) The cross-table tells us that Whites and Blacks do differ in their religious preferences, but it does not do so by ranking those religions, then working math on them. It does so straightforwardly, by showing us the different values in the various table cells.

	White	Black
Prot	52.4% (1171)	74.8% (323)
Catholic	26.0% (580)	7.9% (34)
Jewish	2.7% (61)	0.5% (2)
Other	4.4% (98)	5.1% (22)
None	14.5% (324)	11.8% (51)

The fact is, we can’t use a t-test or ANOVA on a categorical dependent variable. There has to be at least some ordering principle at work, for the results to make any sense. Sex, race, marital status, religion, region, etc. are out of bounds.

We can, however, use ordinal data such as the Socio-Economic Index (Variable 19) as a dependent variable. We can do this because the computer’s numerical score for each of the four social classes (Lower, Working, Middle, Owner/Professional) produces a ranking that makes sense. Though the resulting mean and standard deviation are not as accurate as they would be with interval/ratio data, they are not ludicrous. The resulting t-test or ANOVA does give us a meaningful picture of American social life.

Try it! **Type “16”** as the independent variable and **“19”** as the dependent variable at the “T-Test and ANOVA” variable entry screen. First look at the chart without any control variable, then try controlling for sex (variable 4), race (variable 5, then variable 6), age (variable 2), birth cohort (variable 3), region (variable 8), marital status (variable 22), etc. For the American population as a whole, as well as for each of these categories, “Probability > 99.9% ($p < .001$)”. There is only one chance in 1000 that the difference between the columns is the result of **sampling error** – i.e., a result of the particular sample of people that the GSS interviewed in 2000.

T-tests and ANOVAs only work if the independent variable is categorical and the dependent variable is interval/ratio.

(We can usually fudge and use ordinal for either.)

There is one more limitation on t-tests and ANOVAs: they cannot accept interval/ratio data as an independent variable. This reason ought to be as obvious as the foregoing. Simply put, we can’t compare two or more types of people if there is no clear dividing line between them. Human height is continuous, for example: we find people of every height from less than 4’ to more than 7’. People fall into no natural height groupings, and there are no sharp lines that separate one set of people from another. We can only use t-tests and ANOVAs with such data if we divide people artificially, perhaps by labeling some people “tall” and others “short”, then comparing the people whom we have so labeled. We can similarly compare “rich people” against “poor people” only if we have divided the near-infinite income possibilities into a limited number of manageable groups.

This latter limitation does not much affect Sociological Insights[®], because all of its GSS interval/ratio variables have been divided into ordinal rankings. States data, however, are almost all interval/ratio values, which is why t-tests do not appear on the STates menu.

The important thing to remember is that you have to put the right kind of data in the right place. Categorical or ordinal for the independent variable; interval/ratio or ordinal for the dependent variable. Keep this straight and t-tests/ANOVAs can be a valuable tool.

LOOKING AT DIVORCE

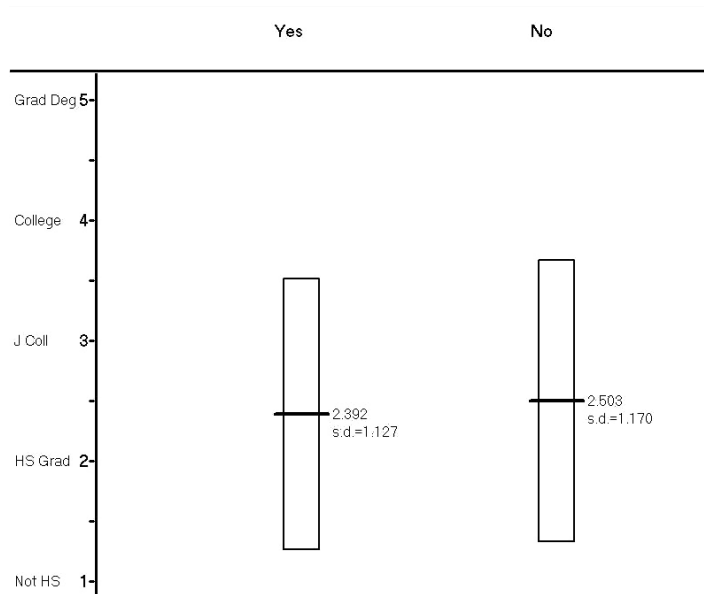
Let’s explore one more topic before we close this chapter. **Choose “Test for Similar Means”** from the **SURvey** menu. **Type “24”** (Ever Divorce?) into the independent variable box.

This GSS question was “Have you ever been divorced or separated?” and it was asked only of those people who had ever been married. 43% of the 2082 people who answered said “yes” and 56% said “no”.*

How do people who have been divorced differ from those who have stayed married? **Type “16”** (DEGREE) into the dependent variable box, skip the control variable, and **click “OK”** to get the graph on the following page.

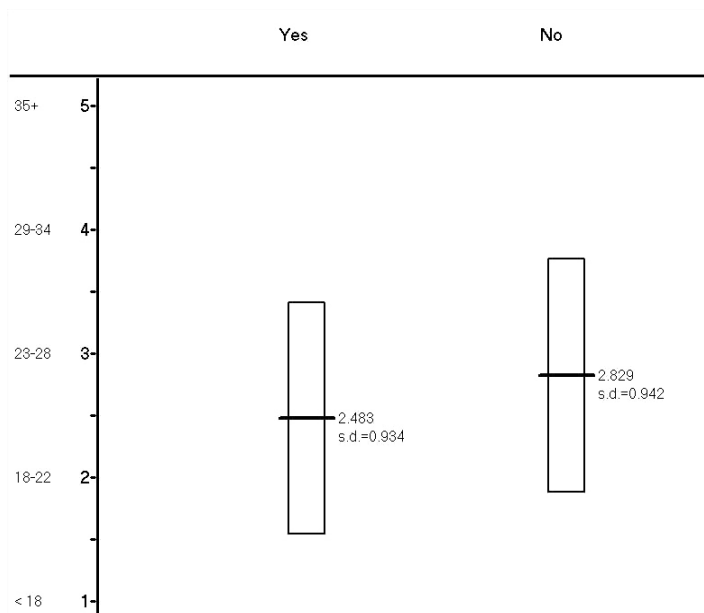
* That 43% includes people who have been divorced multiple times. Perhaps this is the source of the claim that “half of all U.S. marriages end in divorce.” It’s not that half of the people marrying get divorced: it’s half of the marriages that fail, some of which are hopeful attempts by repeat offenders

T-Tests and Analysis of Variance



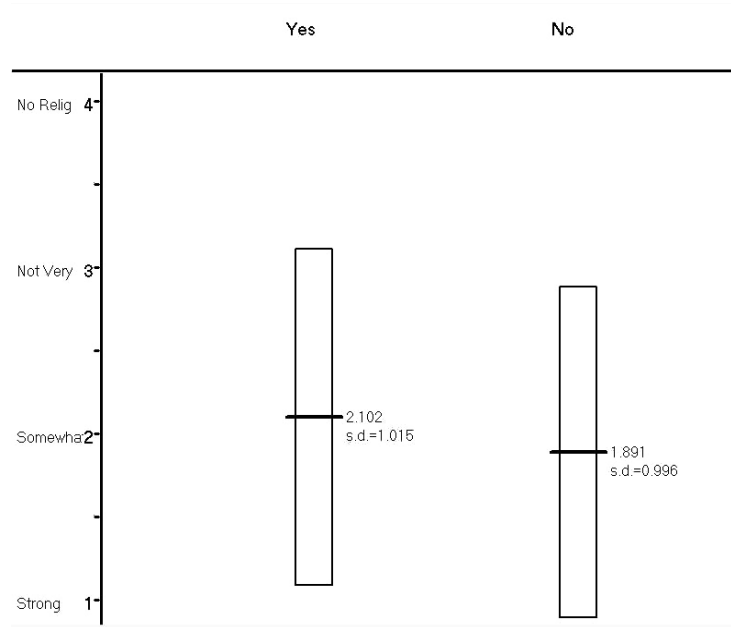
Clearly, with “Probability > 95%” there is an educational difference between these two types of people. Folks who have been divorced (and perhaps remarried) have lower average levels of education than do married people who have stayed married all along. We can’t talk about causality, because the data by themselves don’t tell us whether people get divorced because they are uneducated, or whether they are uneducated because they have to drop out of school and work after they’ve been divorced. But we can note the pattern.

Similarly, if we use **Variable “31”** (Age at the birth of one’s first child) as the dependent variable, we see that people who have been divorced have, on average, begun their families earlier than married people who have never divorced.



Again, the data alone don’t tell us which causes which. It may be that the never-divorced wait until they are more mature to have kids, thus lessening the strain on young marriages. Or it may be because some people with troubled marriages have kids in the hopes that they’ll stay together. The pattern, however, is clear.

Now use **Variable 100** (How religious are you?) as the dependent variable. We see that people who have been divorced tend to be less religious than are others who have married. (In this case, high scores indicate less religiosity; low scores indicate more.)



Do irreligious people divorce more readily, in line with the old saw “the family that prays together stays together”? Or does divorce drive people from their churches, perhaps because a failed marriage shakes their faith? Again, we don’t know. But we see the relationship.

We get similar results comparing **Variable 24** (Ever Divorce?) as the independent variable and using **Variable 13** (Occupational Prestige), **Variable 19** (Socio-Economic Index), and **Variable 47** (Income) as dependent variables. Like the above, all of these divisions give us at least a 95% probability that the difference we see in our GSS sample accurately represents a division in the American population at large. On average, people who have been divorced have lower status jobs, a lower class standing, and less family income than those who have married once and stayed that way.

We’ve already learned quite a lot, but we can do more. Think, for a minute, about the divorce rate among various age-cohorts in the American population. We tend to think of the generation that came of age in the 1940s and 1950s as having more stable lives than we expect today. This generation experienced much more job stability, marital stability, and so on than was the case for either their children (the Baby Boomers) or their children’s children (born after 1970).^{*} And the data show that their married lives, at least, were more stable. The cross-tab of **Variable 3** (Cohort) versus **Variable 24** (Ever Divorced?) is at the top of the next page. (The numbers across the top are birth years.)

^{*} For an accessible account of these changes, read Beth Rubin’s Shifts in the Social Contract: Understanding Change in American Society, Pine Forge Press, 1995.

T-Tests and Analysis of Variance

	->1925	1926-43	1944-57	1958-71	1972+		ROW TOTALS
Yes	23.6% (53)	42.9% (190)	58.9% (371)	39.7% (251)	29.6% (45)		43.7% (910)
No	76.4% (172)	57.1% (253)	41.1% (259)	60.3% (381)	70.4% (107)		56.3% (1172)

As you can see, members of the Baby Boom generation (1944-1957) are much more likely than members of their parents' generation to have been divorced. A clear majority fall into this category, even though many of these may well have remarried successfully. Only a minority of earlier generations ever divorced. The left-most three columns show a clear trend.

At first glance, we might want to interpret the right-most three columns as predicting a declining divorce rate. After all, later birth cohorts have lower levels of divorce than the Boomers. Those born after 1971 have almost as low a divorce rate as do those born before 1926. But we would make a mistake were we to do so, and for two reasons.

First, those in the youngest birth cohort were all under thirty at the time of the 2000 GSS, so there is still plenty of time for their marriages to fail. Experts note that marital dissatisfaction typically increases with the length of a marriage, reaching a peak at the point when a couple's kids are in their teens. Then the kids move out and things improve. Neither of the younger two cohorts has reached that point, so they may have a lot more divorcing to do. We'll have to wait and see.

Second, there has been a major shift in sexual and marriage practices over the last generation – a greater acceptance of sex before marriage and a greater acceptance of people living together unwed. This was still controversial when Baby Boomers were young. It is still controversial for religious conservatives – Southern Baptists, Mormons, and others – but it is not controversial for the majority of the population.* This reduces the divorce rate, because it takes people with less commitment to each other out of the pool.

Still, the rise from the left-most to the middle cohort is quite striking. This is the part of the table that we can use to examine the changing character of divorce in America.

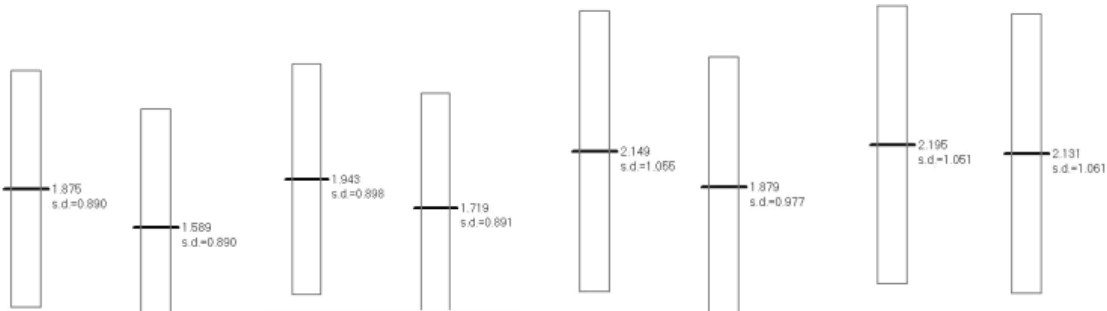
Go back to the **T-Test and ANOVA** variable entry screen. **Type “24”** (Ever Divorce?) as the independent variable, then **type “3”** (Cohort) as the control variable. Successively examine the graphs for **variables “31”** (Age at 1st Kid), **“16”** (Degree), **“13”** (Occupational Prestige), **“19”** (Socio-Economic Index), and **“47”** (Family Income). I don't have room to show the graphs here, but the pattern is pretty much the same for all. Here are the results:

- None of these variables shows any difference for the oldest birth cohort, those people born in or before 1925. Divorced people are just like married people who stayed married in that generation.
- Only family income matters for folks in the 1926-1943 birth cohort. Those in that cohort who have ever been divorced are just like those who stayed married, except that, on average, they get significantly less income.

* Cross-tabulating Variable 97 (as independent) with Variable 177 (as dependent) shows solid majorities of all but Protestant fundamentalists saying that premarital sex is sometimes or always okay. And over 40% of fundamentalists agree. Interestingly, cross-tabulating Variable 97 with Variable 238 shows that similar percentages of religious fundamentalists, moderates, and liberals have had more than one sex partner in the last 12 months. Dogma does not dictate behavior.

- Everything except SEI matters for the Baby Boom cohort. Those who have been divorced, on average, have their kids at a younger age, get less education, have less occupational status, and earn less money than do their non-divorced counterparts.
- Everything matters for those born from 1959 through 1971. People in this cohort who have been divorced have less education, class standing, income, etc.
- Only family income matters for folks born in or after 1972. Members of this cohort who have divorced are more likely to be poor, on average, than are members who have stayed married. As pointed out above, we don't yet know whether this cohort will ultimately have a lower incidence of divorce than its parent cohort did. It is best to leave this group out of our calculations.

We've tested everything except religion: for which birth cohorts does religion make a difference and for which does it not? **Type "24"** into the independent variable box, **type "100"** into the dependent variable box, and **type "3"** into the control variable box. Here are the first four cohorts. (I've left out the youngest cohort because they are too young to have reached their prime divorcing years.)



Of these four cohorts, only the second and third show generalizable differences in religiosity between those who have divorced and those who have stayed married. The probability is greater than 97.5% and 99%, respectively, that the religious differences for the GSS sample of the 1925-1943 and 1944-1957 cohorts reflect a real difference for those cohorts in the whole U.S. population. In both cases, people who have been divorced are, on average, less religious than are those who have stayed married. On the other hand, the probability is less than 95% for the other two cohorts, so we can conclude that, for them, the ever-divorced and the always-married have about the same level of religiosity.

Is there an underlying pattern here? To understand it, take a look at the above graphs. Remember that the higher bars indicate lower levels of religious interest. The most interesting pattern is the steady loss of religiosity on the part of the population as a whole. Older birth cohorts are more religious than are younger birth cohorts – regardless of whether or not they have been divorced. In fact, the Baby Boomer always-marrieds are, on average, a bit less religious than the pre-1925 ever-divorced. This is a clear shift away from religious commitment on the part of the general population.

Now look at the two right-most columns: the 1944-1957 and 1958-1971 cohorts. Compare the ever-divorced bars, and you'll find that they are almost identical. The later cohort is a smidgeon higher in its mean and a smidgeon smaller in its standard deviation, but we're talking decimal points. There has been no real shift in religiosity among the ever-divorced between those two groups. Now compare the two always-married bars, which are significantly different. Those who have stayed married in the younger cohort are much less religious than are their Boomer counterparts.

Let's pull this together into an overall trend. First, there is a steady decline in average levels of self-reported religiosity – a decline that began and ended one cohort earlier among those who have divorced than it did among those who have stayed married. Both ever-divorced and always-married were religious for those born before 1926. The ever-divorced became less religious in the 1926-1943 birth cohort, and continued that shift into the Boomer years. The always-married became less religious beginning with the Boomers, and caught up to the ever-divorced one cohort later. The “religiosity gap” between divorced and un-divorced has shrunk for the post-Boomer cohorts, largely because the un-divorced in those cohorts are not nearly as religious as were those who came before. No gap, then gap, then no gap. Our story is no longer just about divorce, but about wider social trends.

See how much one can learn if one examines things closely?

THINGS TO REMEMBER

1. **T-tests** and **analysis of variance (ANOVA)** measure the significance of the differences between various categories of people on some measure. The people might be in different demographic categories, might be experimental subjects in a medical trial who either took a drug or a placebo, etc. The measure can be anything from which one can figure a reasonable mean and standard deviation. This means that:
 - Independent variables are ideally **categorical**, but may be **ordinal** if it is their differences rather than their rank order that matters.
 - Dependent variables are ideally **interval/ratio**, but may be **ordinal** if there are enough categories and they are evenly enough distributed for it to make sense to talk about their **measures of central tendency** and **dispersion**.
 - The computer will always let you forget this. You have to think about what you are doing. If you are trying to take an average of something like race or sex, you are on the wrong track.
2. **T-tests** are used with variables that have two categories. **ANOVAs** compare more than two categories. (An ANOVA will tell you whether there are significant differences between the columns as a whole without telling you exactly where that difference is. A more complicated program than Sociological Insights[®] will let you use a t-test to compare any two categories once you've discovered the wider pattern.)
3. Both tests can be used with much lower numbers of people than are required for the **chi-square** test. Usually, one needs only about 30 participants to draw meaningful conclusions.
4. One reads the “Probability” and “p” figures exactly the same way that one reads them for the chi-square test. “Probability” must be equal to or greater than 95% (which makes p less than or equal to .05), in order to conclude that the differences we see in the sample reflect actual differences in the wider population.
5. One uses categorical control variables with both t-tests and ANOVAs, in exactly the same way that one uses them with cross-tabulations.

TEN: STANDARDIZED CROSS-TABULATION

At the start of the last chapter, we identified a problem with ordinary cross-tabulations. Too often, a cross-table lacks enough data in each cell to test the significance of the differences between its columns. T-tests and ANOVAs are one solution to this, but they only work when the dependent variable is interval/ratio or ordinal data. If the dependent variable is categorical data, they don't do much good.

Let's look at an example. Let's say that we are trying to trace the relationship between the kinds of families in which people were raised (Variable 36) and their current marital status (Variable 22). An ordinary cross-tabulation shows us that people raised in single-mother households or with a mother and step-father are more likely to be divorced or never married than are people raised by both parents, by single dads, or by dads and step-moms. (Scan that sentence again: like much sociology, it's complicated – but it's very precise.)

We know that current marital status varies with the region in which people were raised (Variable 34). People raised in the West and Northeast are much more likely never to have married than people raised in the Midwest and South. “Probability > 99.9%” for each of these cross-tables, so the patterns are very sure. Here they are:

	Both	Dad/Step	Mom/Step	Sing Dad	Sing Mom		NthEast	South	Midwest	West
Married	50.0% (968)	42.6% (26)	34.4% (52)	47.6% (90)	32.1% (134)	Married	43.1% (258)	45.0% (401)	47.1% (336)	42.9% (181)
Widow	10.2% (197)	14.8% (9)	7.3% (11)	11.1% (7)	5.0% (21)	Widow	11.4% (68)	10.9% (97)	9.8% (70)	6.6% (28)
Divorce	14.3% (276)	13.1% (8)	18.5% (28)	9.5% (6)	17.7% (74)	Divorce	12.0% (72)	16.7% (149)	16.1% (115)	16.1% (68)
Sep	3.1% (60)	3.3% (2)	5.3% (8)	4.8% (3)	7.2% (30)	Sep	4.5% (27)	5.3% (47)	2.1% (15)	3.6% (15)
Nev Mar	22.4% (434)	26.2% (16)	34.4% (52)	27.0% (17)	37.9% (158)	Nev Mar	29.0% (174)	22.1% (197)	24.8% (177)	30.8% (130)

The question, though, is how these patterns interact. Is the region in which one was raised an independent influence on one's current marital status – perhaps because people in different regions have different cultural expectations for family life? Or does regional influence operate through the medium of childhood family structure? It may be, for example, that the West's large proportion of never-married people is the result of that region's relative lack of two-parent families and greater percentage of single-mother households.

In fact, Variables 34 and 36 are also highly related. All three variables are very much tied up with each other. The question is, What are their real relationships?

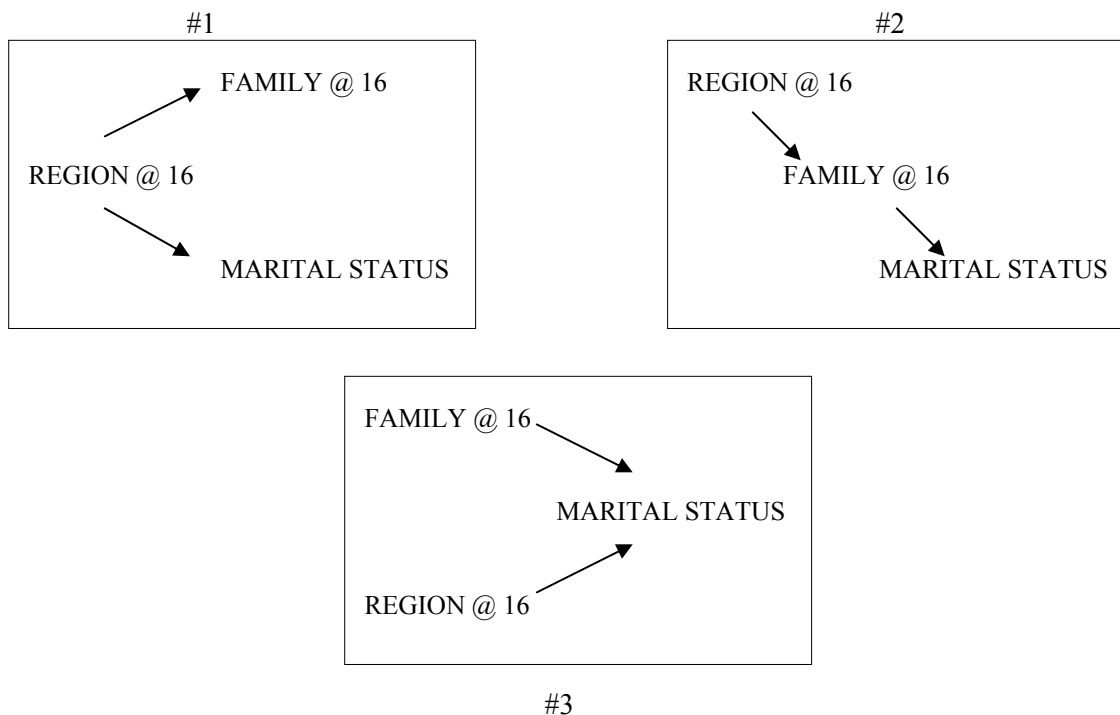
The issue is much like the one we explored in Chapters 4 and 5, in which we discovered that some apparent correlations between variables were spurious and others were not. In this case, we have three possibilities:

1. Region of origin influences both the kind of family that one grew up in and one's current marital status, but there is no other direct connection between these latter two. (I.e., the relationship between natal family structure and current marital status – in the leftmost of the two tables above – is **spurious**, because it is an artifact of both family structures being independently related to region.)

Standardized Cross-Tabulation

2. Region of origin influences the kind of family one grows up in, and this natal family structure influences one's current marital status, but there is no connection between region and current status *per se*. (I.e., the rightmost of the above tables is **spurious**.)*
3. Region of origin and natal family structure both influence people's current marital status, and they do so independently. (I.e., neither relationship is spurious.)

Diagrammatically, we have:



Normally, we would use a controlled cross-tab to test these relationships. Keeping Variable 22 (Marital Status) as the dependent variable, we would take the following steps:

1. We first would use Variable 36 (Family @ 16) the independent and Variable 34 (Region @ 16) the control.. If the relationship between the kind of family one grew up in and one's current marital status disappeared, then we would know that Box #1 accurately describes the situation.
2. We would then reverse the independent and control variables, making Variable 34 (Region @ 16) the independent variable and Variable 36 (Family @ 16) the control variable. If the relationship between the region in which one was raised and one's current marital status disappeared, we would know that box #2 accurately describes the situation.
3. If neither relationship disappears, then we know that Box #3 is right.

Try it! Use **Variable 22** (Marital Status) as the dependent variable, and (alternately) **Variable 34** (Region @ 16) and **Variable 36** (Family @ 16) as the independent and control variables. Try to tell whether the patterns are significant.

* It does not make much sense to pose the opposite case: that natal family structure "causes" region, which then causes current marital status. This would require that people with particular family structures move to a congenial region. We can eliminate this hypothesis from discussion .

Oops! You can't! "Probability = N/A" in all but two of the many possible cross-tables. The General Social Survey sample is not big enough to fill this complex a table's cells with enough people to figure the chi-square test of significance. Too few people in the sample were raised by single dads or by one parent and one step-parent. We can't carry out the analysis that we need.

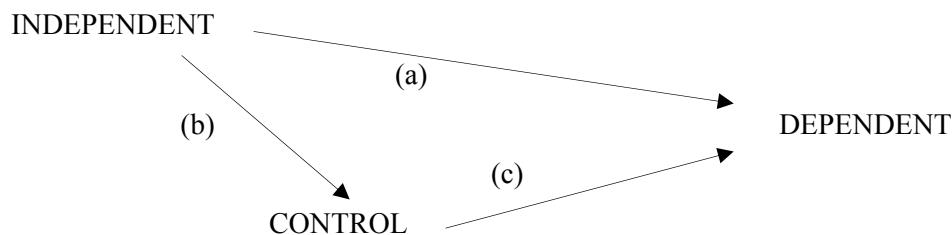
We also can't use t-tests or ANOVAs, because all three variables are categorical. These tests require that the dependent variable be interval/ratio (or at least ordinal), at that is not the case here.

ANOTHER METHOD

There is, of course, another method available to us. It is rarely used by psychologists and medical researchers, and is thus not found in most elementary statistics texts. But it works wonderfully with sociological survey data. It makes it possible for sociologists to trace relationships that do not lend themselves either to ordinary cross-tabulations or to t-tests/ANOVAs.

The technique is called **standardized cross-tabulations**. Developed by Morris Rosenberg and extended by James A. Davis,^{*} it works much like multiple regression. Sociological Insights[®] begins with cross-tables between all the variables, and progressively eliminates the interactions between the independent variables. The result is a clear picture of the effect of an independent variable on the dependent variable, with the influence of other independent variables removed.

Take the following **path diagram**, which charts the possible lines of influence between an independent variable, a dependent variable, and a control variable. The independent variable can work directly on the dependent variable (path "a"). Or it can work indirectly through the control variable, by means of paths "b" and "c". Standardized cross-tabulation removes the path "b-c" from consideration. It asks whether the independent variable influences the dependent variable directly.



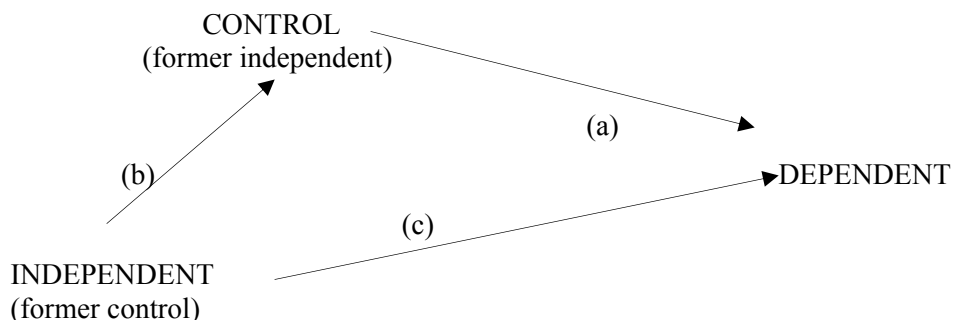
If our standardized cross-tabulation shows no difference between the columns of the independent variable, once path "b-c" has been removed, then we know that the relationship between the independent variable and the dependent variable is spurious. In such a case, path "a" does not exist and the true lines of influence are like those in box #2 on page 112. If, however, there is a significant difference between the columns of the independent variable, then we know that path "a" is a way that the independent variable has influence. Either box #1 or box #3 is a possibility.

Unlike multiple regression, standardizing cross-tabulations only works on one independent variable at a time. That is, it tests whether the independent variable influences the dependent

^{*} See Morris Rosenberg, The Logic of Survey Analysis, New York: Basic Books, 1968 and James A. Davis, "Extending Rosenberg's Technique for Standardizing Percentage Tables. Social Forces, 62: 679-708, 1984. Davis provides many extended sociological applications in his now-out-of-print Social Differences in Contemporary America, New York: Harcourt Brace Jovanovich, 1987.

variable, after the indirect path through the control variable (“b-c”) has been eliminated. We can produce another table that measures the strength of line “c”, but this would be the same as a standard cross-table because our model does not suppose that Independent Variable #2 influences the Dependent Variable via lines “b” and “a”. To test that, we need a separate diagram:

To produce this chart, we enter our former control variable as the independent variable, then our former independent variable as the control variable. The resulting cross-table shows us the influence of the former control variable on the dependent variable (path “c”), after the influence through the former independent variable (path “b-a”) has been removed.



Again, if there is a difference between the columns of the standardized cross-tabulation after path “b-a” has been eliminated, then path “c” is a true path of influence. If there is no difference between the columns, then path “c” is spurious. By systematically testing each of these paths of influence, we can arrive at a final picture of the relationship between these variables. With three variables, this will one of the paths diagrams show in Box #1, Box #2, or Box #3 on page 112.

This technique has two advantages over ordinary controlled cross-tabulations. First, it eliminates the problem of small cell numbers. Unlike ordinary cross-tabulations, standardized cross-tabulation does not subdivide the sample into smaller and smaller boxes. Instead, it adjusts the data to expose the real relationships while maintaining the same (or nearly the same) sample size.

Second, it allows us to enter several control variables. I won’t do that in this chapter, because I want to keep the examples simple – though we will test a more complicated path diagram than this three-variable one. Just remember that it is possible to work with up to five variables simultaneously.

Standardized cross-tabulations work with categorical or ordinal data.

Let’s apply this technique learn about the relative influence of region-of-origin and natal family structure on marital status.

FAMILY OF ORIGIN? REGION? OR BOTH?

Start Sociological Insights[®]. Click on the SURvey menu, then choose “Standardize Cross-Tabulation”. Use Variable 36 (Family @ 16) as the independent variable, Variable 22 (Marital Status) as the dependent variable, and Variable 34 (Region @ 16) as the control variable. Click “OK” to show screen at the top of the next page.

Standardized Cross-Tabulations

Graphs Remove Controls Restore Controls Print Screen Get Help Close

Relationship Between Family @16 and Marital after Controlling for Region @16

Show: Family@16 → Marital Show: Region@16 → Marital

Family @16: With whom did you live at age 16? (Other = missing data)

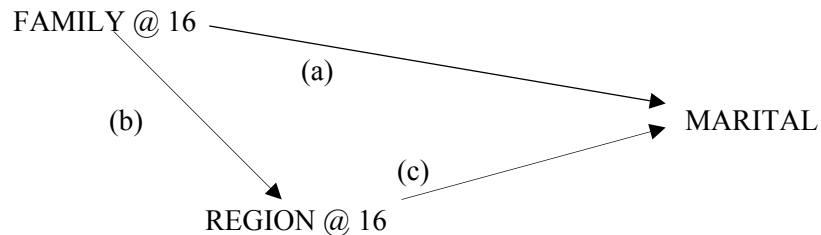
	Both	Dad/Step	Mom/Step	Sing Dad	Sing Mom		ROW TOTALS
Married	49.2% (894)	43.5% (77)	33.1% (49)	47.5% (29)	31.8% (123)		45.3% (1112)
Widow	10.5% (188)	14.5% (8)	9.5% (14)	11.5% (7)	5.4% (21)		9.7% (236)
Divorce	14.8% (252)	11.3% (7)	17.6% (26)	9.8% (6)	17.3% (67)		14.6% (368)
Sep	3.3% (59)	4.8% (3)	5.4% (8)	4.3% (3)	6.7% (26)		4.0% (99)
New Mar	23.8% (413)	25.8% (16)	34.5% (51)	26.2% (16)	38.8% (150)		26.3% (646)
COLUMN TOTALS	100% (1796)	100% (82)	100% (148)	100% (81)	100% (387)		

2454 interviewees

Probability > 99.9% (p < .001)

Let's walk around this screen a bit. The cross-table looks familiar, as indeed it is: it shows the relationship between the independent variable (Family @ 16) and the dependent variable (Marital Status).

The title at the top, however, tells us that the table shows this relationship after controlling for the influence of "Region @ 16" on the dependent variable. We are seeing the net effect that one's family of origin has on one's marital status, after the effect of region situation has been removed. We are seeing line "a" in our path diagram:



There is a menu bar at the top left corner of the screen. The second menu item from the left says "Remove Controls". **Click that menu item** and watch the table change. This shows us the relationship between Family and Marital Status, with no control variable. We say this table on page 111, on the left – the ordinary cross-tabulation between these variables. We can use "Remove Controls" and "Restore Controls" to switch between the two tables. Clicking "Remove" shows us the total influence of our independent variable on the dependent (path "a" plus path "b-c"). Clicking "Restore" gives us the net influence of that independent variable, after eliminating any influence that operates through the control (path "a" alone).

In this case, there is no difference. "Probability > 99.9%" for both the raw cross-table and for the controlled one. This tells us that the structure of one's family of origin is a completely independent predictor of people's current marital status. Path "a" has real influence, and that influence is not increased by influence along path "b-c".

Standardized Cross-Tabulation

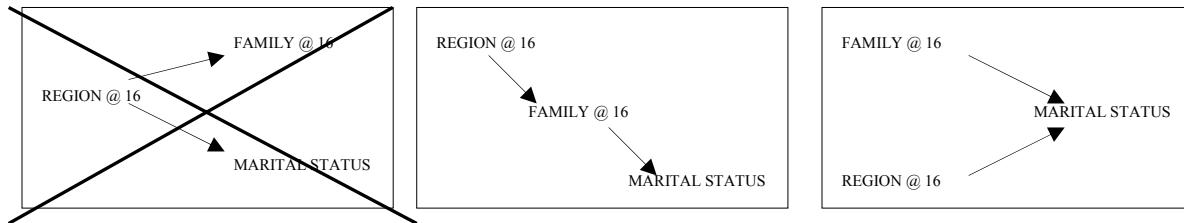
We can thus eliminate diagram #1 on page 112 from among the possible relationships between these three variables. Family of origin does not influence current marital status just because of its relationship with one's natal region. It influences current marital status on its own.

In a moment, we'll examine which of the remaining diagrams (#2 or #3) on page 112 best depicts the relationship between these variables. For now, we need to see what else this screen can tell us.

Take a look at the row of buttons just below the label "Relationship Between Family @ 16 and Marital after Controlling for Region @ 16". The one on the left is blanked out, so **click on the one on the right**. This presents a new table, titled "Relationship between Region @ 16 and Marital (no controls)". This is the same table that we saw on page 111, on the right – the raw relationship between these two variables, with no controls. This stands for line "c" in the path diagram. Were we working with more than one control variable, we would have a whole line of buttons to choose from. As we are not, we just have two.

The "graphs" menu item works just like it did with ordinary cross-tabulations. Sociological Insights[®] can give us both telegraph bars and column distribution charts. These sometimes help us see relationships more easily than do raw numbers alone.

Here are our three diagrams again: #1 on the left, #2 in the middle, #3 on the right.



So far, we have eliminated diagram #1. (I have crossed it out.) Let's work on the other two.

Close this screen, and **type "34"** into the "Independent" box on the variable entry screen. Leave the dependent variable just as it is, and **type "36"** into the "Control #1" box. **Click "OK"** to show the following table:

Standardized Cross-Tabulation

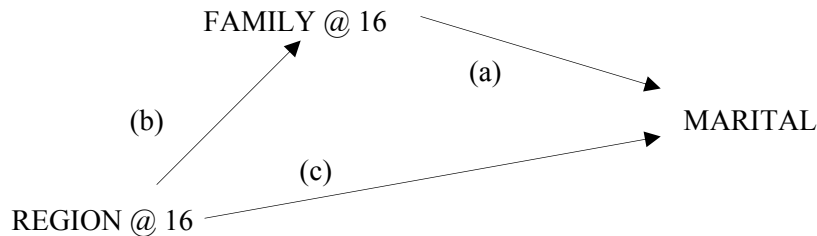
Relationship Between Region @ 16 and Marital after Controlling for Family @ 16

Region @ 16: Region lived in at age 16

	NEast	South	Midwest	West			ROW TOTALS
Marital: Current marital status							
Married	45.8% (251)	48.3% (322)	50.7% (348)	45.5% (191)			48.0% (1174)
Widow	11.8% (64)	11.3% (82)	9.8% (66)	7.8% (31)			10.4% (229)
Divorce	11.6% (64)	15.3% (134)	14.6% (98)	14.1% (56)			14.1% (344)
Sep	5.1% (29)	4.1% (30)	1.6% (11)	3.8% (15)			3.5% (89)
Never Mar	25.6% (140)	21.8% (170)	23.2% (159)	29.5% (117)			24.0% (568)
COLUMN TOTALS	100% (567)	100% (811)	100% (671)	100% (337)			

2446 interviews
Probability = > 95.5% (p = < .005)

This screen is just like the other one, except that we have reversed the independent and control variables. This time, we are looking at the “Relationship between Region @ 16 and Marital after Controlling for Family @ 16”. The question before us is whether the region in which one was raised influences one’s marital status directly (path “c”), or only because family structure varies by region (path “b-a”). If the latter is the case, then we know that diagram #2 is right. If the former is the case, then we opt for diagram #3.



Just as before, the key is whether there is a significant difference between the columns of the table we are viewing. A glance to the bottom of the screen shows us that the “Probability > 99.5%” that these differences are significant. One’s region of origin therefore does have an independent direct influence on one’s current marital status. We can eliminate diagram #2 from consideration. We are thus left with diagram #3, which shows that both region and natal family independently influence one’s current marital status.

There is, however, one more subtlety to explore. **Click the “Remove Controls” button.** The “Probability” jumps from 99.5% to 99.9%. This isn’t much, but it tells us that the direct influence of natal region on marital status (path “c”) is just a bit lower than its total influence (path “c” PLUS path “b-a”). Some of the influence that region has on marital status therefore comes from region’s influence on family structure. Most of it flows directly (path “c”) but some also flows along path “b-a”.

Summarizing all this, we get the following, which is illustrated by the diagram at the top of this page:

- Children raised by single mothers and by mothers who have remarried are significantly more likely to be separated or divorced than are those raised in other circumstances. And they are also significantly more likely to remain single. (This is path “a” in the diagram above.)
- Simultaneously, people raised in the Midwest and South are significantly more likely to have married than have people from other regions. People raised in the West are least likely to have been married, and people raised in the Northeast are least likely to have divorced. (This is path “c” in the diagram.)
- People raised in the West are least likely to have been raised by both parents and are most likely to have been raised by single parents. (This is path “b” in the diagram.)

In the final analysis, both factors influence patterns of marital status, independently.

WEALTH AND HAPPINESS

Let’s now take a more complicated example. In Chapter 6, we took a quick look at what kinds of people are happy and what kinds of people are not. Among other things, we saw that men and women are equally happy and that fewer African-Americans than Whites consider themselves

“Very Happy”. Here are a few more cross-tables to consider:

Family Income (Prob > 99.9%)

	< \$15K	\$15-34K	\$35-59K	\$60-109K	\$110K+
Very	19.7% (95)	26.8% (195)	33.8% (202)	42.5% (188)	45.6% (78)
Pretty	58.1% (280)	61.5% (448)	60.2% (360)	59.2% (235)	49.7% (85)
Not Too	22.2% (107)	11.7% (85)	6.0% (36)	4.3% (19)	4.7% (8)

SEI (Prob > 99.9%)

	Low	Working	Middle	Prof/Own
Very	25.5% (121)	29.0% (239)	34.3% (287)	39.3% (190)
Pretty	59.1% (280)	60.8% (501)	56.7% (474)	53.2% (257)
Not Too	15.4% (73)	10.2% (84)	9.0% (75)	7.5% (36)

Education (Prob. > 99.9%)

	Not HS	HS Grad	J Coll	College	Grad Deg
Very	25.6% (109)	29.6% (440)	35.0% (71)	38.7% (167)	42.0% (89)
Pretty	54.6% (232)	60.3% (897)	56.2% (114)	55.3% (239)	51.9% (110)
Not Too	19.8% (84)	10.1% (150)	8.9% (18)	6.0% (26)	6.1% (13)

Occupational Prestige (Prob > 99.9%)

	<35	36-45	46-55	56-65	66+
Very	26.5% (229)	29.4% (165)	36.0% (231)	38.7% (129)	37.5% (84)
Pretty	60.4% (521)	59.1% (332)	56.5% (363)	54.1% (180)	54.0% (121)
Not Too	13.1% (113)	11.6% (65)	7.5% (48)	7.2% (24)	8.5% (19)

Each of these independent variables measures social status. In each case, a larger percentage of those with greater status say they are “Very Happy”. If these figures are correct – and the “Probability” numbers tell us that there is only one chance in 1000 that they are not – greater income, higher social class standing, more education, and higher occupational prestige all predict greater happiness.

We know, however, that greater income, etc. all go together. People with lots of education tend to have high prestige jobs, which brings them lots of money and locates them higher in the class system. Do all of these factors cause happiness independently? Or do some of them only work through others.

Just like with region of origin, family of origin, and marital status we have several possibilities. Here, we’ll focus on two. Either:

1. All these variables independently influence people’s happiness. (This is like diagram #3 on page 112.)
2. At least some of the variables work through others but do not affect happiness directly. “A” may influence “B”, which in turn influences “C”, even though “A” does not influence “C” directly. (This is like diagram #2 on page 112.)

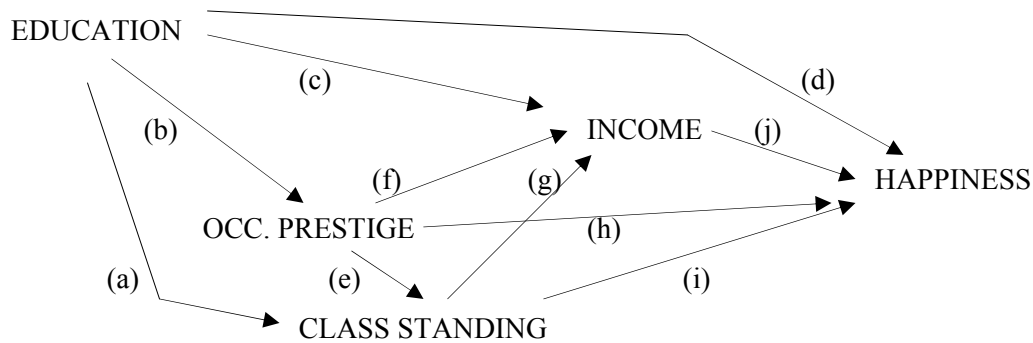
Standardized cross-tabulations let us determine which of these is true.

Before we start, we should do a little thinking. What might be the causal relationship between the various factors that we have identified here?

- Clearly, education helps people get high prestige jobs. This is not true for sports stars, but it is for pretty much everyone else.
- Generally speaking, high prestige jobs usually bring in more income than low prestige ones. There are exceptions – clergy and teachers, to give two examples – but they are relatively rare.
- Job prestige pretty much matches class standing. In fact, Variable 19 (SEI) uses employment level as a measure of class, so we would expect class standing and occupational prestige to be pretty much the same thing.

- One’s parents’ socio-economic standing helps determine what kind of education one gets, and thus where one ends up in the class structure. But education usually comes first in determining one’s own status.
- Money can buy education, but that usually happens after one is too old for that education to make much difference in one’s life.
- People often say “money can’t buy happiness”, but you don’t often see them giving it away on the street corners. If folklore is any guide, the relationship between money and happiness seems pretty direct.

Here’s a **path diagram** that summarizes all this. Let’s start by seeing which of the following lines of influence (a, b, c, d, e, f, g, h, i, j) are important and which are spurious:



Choose “**Standardize Cross-Tabulations**” from the SURvey menu, then **type “19”** (SEI) into the independent variable box. **Type “113”** (Happy?) into the dependent variable box, and **type “47”** (Family Income) into the first control variable box. **Click “OK”** to get the following diagram.

	Low	Working	Middle	Prof/Own
Very	27.0% (110)	29.3% (213)	30.9% (228)	33.0% (140)
Pretty	61.9% (252)	61.2% (444)	58.9% (435)	55.0% (233)
Not Too	11.1% (45)	9.5% (69)	10.2% (75)	12.0% (51)

With “Probability = 50-74%”, there is clearly no significant relationship between class standing and happiness, once we have controlled for family income. **Click the “Remove Controls” menu item**, just to remind yourself that there was a significant relationship before we controlled for income. Any influence that SEI has on happiness stems from the influence of class standing on income. **Cross out path “i” on the diagram above**. That line of influence is spurious.

Close this standardized cross-table, then **type “13”** (Occupational Prestige) into the independent variable box. Leave the other variables the same, then **click “OK”** to get chart at the top of the next page.

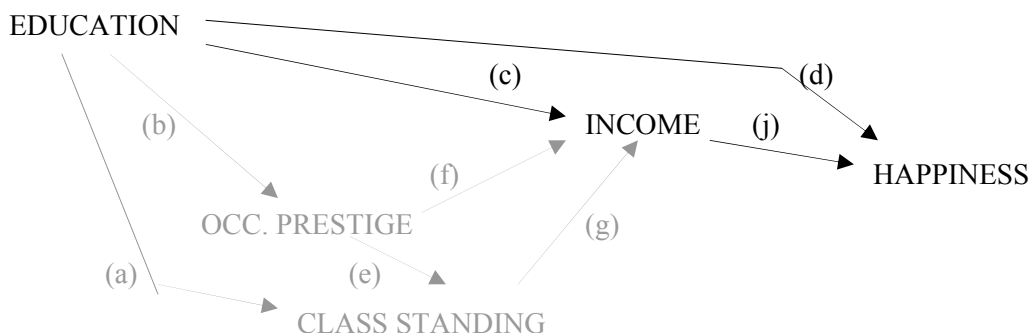
Standardized Cross-Tabulation

	<35	36-45	46-55	56-65	66+
Very	29.0% (220)	29.8% (145)	33.0% (186)	35.9% (107)	30.6% (60)
Pretty	60.2% (457)	60.2% (293)	58.0% (327)	53.7% (160)	57.1% (112)
Not Too	10.8% (82)	10.1% (49)	9.0% (51)	10.4% (31)	12.2% (24)

Here, too, “Probability = 50-74%”. There is no significant relationship between occupational prestige and happiness, once we have controlled for family income. Any influence that occupational prestige has on happiness stems from the influence of class standing on income. **Cross out path “h” on the diagram on page 119.** That line of influence is also spurious.

Now run the same analysis using Variable 16 (Degree) as the independent variable. This time, there remains a significant difference between the columns, even after controlling for income. “Probability > 95%”, which is high enough to be sure that the rise in the percentage of people claiming to be “very happy” as we move from left to right is a real pattern in the population as a whole. We can justifiably leave path “d” in the diagram.

Education is independent of family income in predicting happiness, but is the reverse true? To find out, we reverse the positions of these two variables. Use “47” (Income) as the independent variable and use “16” (Degree) as the control variable, keeping “113” (Happiness) in the dependent position. “Probability > 99.9%”, so income and education both contribute to happiness, independently of each other. Our four possible independent predictors of happiness have been reduced to two. Both paths “d” and “j” are real, showing that education and income influence people’s reported happiness. The other variables that we tested do not.



Note that there could still be relationships between the various antecedent variables (paths a, b, e, f, and g). I have printed them in gray in this diagram because they do not influence happiness – and are thus not relevant to the question we are investigating. Occupational prestige could influence income, as could class standing. And we will see in a moment that education certainly does. But only education and family income influence people’s reported happiness directly.

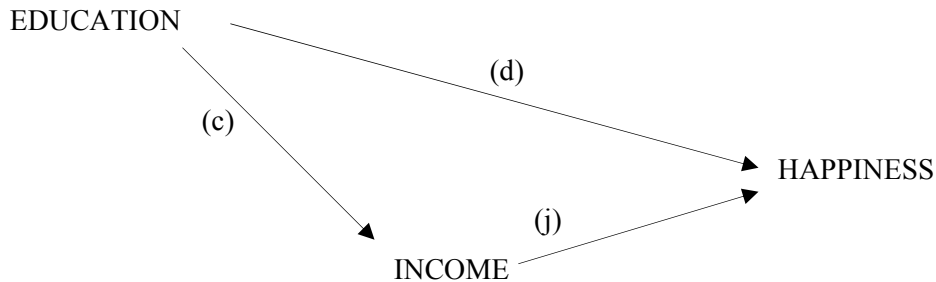
We still need to investigate path “c”. That path (between education and income) implies that education promotes happiness not just by itself (path “d”), but also by increasing income. More schooling produces more income, which in turn produces more happiness. And more

education produces more happiness all by itself. Do both paths work? Or is path “d” the only line by which education influences happiness?

We test this by comparing the raw relationship between education and happiness with that same relationship after controlling for family income. Enter “16” (Degree) as the independent variable and “47” as the control variable. When the cross-table appears, **click “Remove Controls”**. We see that the raw probability was “> 99.9%”. Click “Restore Controls” and it drops back to “> 95%”. This tells us that some of education’s influence on happiness flows through its effect on income – via path “c”. Most of it, however, flows through path “d”, producing happiness directly..

Does family income also influence happiness through education? Or is all of its influence direct. Use “47” as the independent variable again, and **type “16”** in the first control box. In this case, the raw relationship between Income and Happiness is identical to its relationship after controlling for education. “Probability > 99.9%” in both cases. This means that income influences happiness directly, without any portion of its influence flowing through education.

Here is the final diagram, with all of the variables that don’t influence happiness removed:



As you can see, this process is rather complicated. It takes a lot of thinking to develop a path diagram, and it takes a lot of testing to see which paths are real sources of influence between the variables.

At this point, you may want to explore the influence of occupational prestige and class standing on income (paths “f” and “g” in the diagram on the previous page), along with the influence of education on each of these variables (paths “a” and “b”) and the influence of occupational prestige on class standing (path “e”). None of these relates to happiness, but finding out which factors influence people’s incomes is a worthy sociological goal. Having worked through a study of some of the social factors that influence happiness, you should have no trouble with this new task.

MULTIPLE CONTROL VARIABLES

Neither of the examples in this chapter needed more than one control variable. As you have noticed, however, *Sociological Insights*[®] allows you to use up to three control variables in a standardized cross-tabulation. I shall not provide an full example that does so, because I am sure that nobody wants to read a 20-page chapter. I shall, however, talk through a modification of our last example, to show you how it is done. The logic is really rather simple.

Let’s start with the fact that race and income are both independent predictors of people’s happiness. (You can check this yourself, using either of our previous examples as a model.) If we want to learn whether education also has an independent influence on happiness, separate from both race and income, we would have to control for both in our cross-tabulation. To do this, we run a standardized cross-tabulation using Variable 16: DEGREE as the independent variable,

Variable 113: HAPPY? as the dependent variable, Variable 6: RACE B/W as the first control variable and Variable 47: FAM INCOME as the second control variable. Examining the difference between the columns, we find that the “Probability > 99.9%” that education influences happiness, independently of both race and income.

Clicking the “Show Race B/W -> Happy?” button proves that race influences happiness after controlling for income. Removing the controls raises the Probability from 97.5% to 99.9%, so we know that some of that influence is indirect: race influences income, which in turn influences happiness. But race also influences happiness directly.

If we can make race the independent variable and education the first control variable, we learn that race predicts happiness independently of both education and income. We cannot explain African Americans’ lower levels of happiness, compared to Whites, by their lower level of education and income. Race matters, too.

As I said, the process is not hard. Now that you know how to do it, you can explore some very complex social patterns.

THINGS TO REMEMBER

1. **Standardized cross-tabulations** work just like normal cross-tabulations. The only difference is that the values in each cell are adjusted to highlight the direct relationship between the independent and dependent variables, eliminating the influence of one or more control variables.
2. Sociological Insights[®] returns a chi-square test of significance for each table. Read this just as you read the chi-square tests for normal cross-tabulations.
 - Standardizing cross-tabulations eliminates the problem of small cell sizes, which makes it much more likely that the chi-square statistic will work.
 - It also makes it possible to control for several variables simultaneously.
3. Order matters! Enter the independent variable, then the dependent variable, then the various control variables in the order in which you think they might influence each other. It usually helps to draw a **path diagram** to make this clear.
4. Test each link on the various paths to see which ones are real connections and which are spurious.

EPILOGUE: DOING SOCIOLOGY

By now, you have a pretty good idea of how to do quantitative sociology. Not only has each of the preceding chapters shown you one or more ways to make sense of numerical data. Each chapter has also used these techniques to explore one or more sociological questions.

By working through the examples, you have learned how to think sociologically. That is the point of quantitative sociology, after all: to use numbers, trends, and so on, to think through sociological issues. You are now ready to strike out on your own.

Below, I have listed some topics that you can explore using Sociological Insights[®] and its data sets. I am sure that you or your instructor can think of others. Enjoy!

SUGGESTED PROJECTS

1. Exploring Crime Rates

We explored crime rates in several of the previous chapters, most notably looking at the changes in the distribution of burglary rates over the last 65 years in Chapter 2. We saw that the explosive rise in burglary rates in the 1980s was concentrated in just a few states. We also saw that the decline in burglary rates in the 1990s brought those states back in line with the rest.

Variables 68-70, 98-120, 167, and 218-243 of the States dataset contain data on various other crimes and crime rates. GSS variables 58, 188-194, and 246-247 all have something to do with crime. Here are some ideas about how to use them:

- Examine the changes in murder rates, assault rates, rapes, suicides, and rates of alcoholism from the 1920s to the 1990s. In what regions were each of these matters concentrated in the various decades for which we have data? Use the “Show Statistics” menu item on the “Map Variable” screen to display the means, medians, ranges, and standard deviations of these variables. How have these changed from decade to decade? How have the bar charts shifted? What does this tell you about the underlying social trends?
- Determine which factors correlate with violent crimes. To the extent that the data allow you to do so, determine which of these factors correlate through the decades and what shifts (if any) have occurred. Which of these factors are apt to be causal? Which are apt to be spurious? Why?
- Determine which factors correlate with property crimes. Check these across the decades, looking for regularities. What causal connections can you locate? Which connections are spurious?
- Based on the patterns of location that you explored above, plus the patterns of cross-decade correlation, does the distinction between “violent crimes” and “property crimes” make sense? Are they different enough to be seen as separate phenomena? Which crimes fall into each category? Which crimes don’t fit? How would you categorize rape? Suicide? Alcoholism? Do we need more categories of crime than these two? If so, how many others and what might we call them?
- Use GSS data to determine what demographic factors accompany various attitudes toward crime. How do age, race, education, social class, gender, political

allegiance, and other factors influence – for example – people’s attitudes toward police violence? Towards drug use?

- What other attitudes seem to accompany various people’s attitudes toward crime? Are people who are “soft” on crime more sexist? Or less? More likely to oppose abortion? Or less? Are there any constellations of attitudes that seem to go together? If so, what are they? Can you make any demographic sense of such constellations?

2. Residential and Regional Mobility

Chapter 5 underlined the social importance of residential mobility. Among other things, we showed that suicide rates are higher where there is more residential mobility – independent of other factors. Regional mobility also matters. People who grew up in the South tend to have different attitudes than people who grew up elsewhere. You can explore these factors using both States and GSS data. For example:

- What impact does residential mobility have on other aspects of social life? What things besides suicide happen in places where people move around a lot? What do you think might cause them, other than residential mobility – and the lack of social ties that it implies? (Use Chapter 5’s analysis of suicide as a model to help you look at the interaction of several factors.)
- States variables 5-7, 28, 122, 140, 215, 248, 250, 253, etc. all measure some aspect of population movement. Examine the patterns that you found above using some of these other measures of mobility. Which are the same? Which are different? Why might this be the case?
- Explore residential mobility using GSS data.
 - Variable 8 lists the respondent’s current region, and variable 34 lists the region where the respondent lived at age 16.
 - Variable 9 tells whether the respondent currently lives in a big city, suburb, town, or rural area; variable 33 tells in which of these she or he lived at age 16.

What attitudes differ, if any, between those who now live in the same place they grew up and those who moved? How do these attitudes differ between migrants from various regions? Between those who moved from city to town and those who moved from town to city?

(You will need to use the present region or place as the independent variable and the region or place at 16 as a control variable. Compare across both columns and screens. Use controlled cross-tabulations, t-tests/ANOVAs, and standardized cross-tabulations, as appropriate.)

3. Exploring Divorce

In Chapter 8, we explored various aspects of divorce, using GSS data. Among other things, we showed that people who had been divorced – even if they had subsequently remarried -- were more likely to be less religious, to be less educated, to have become parents earlier, and to have lower occupational prestige and socio-economic status than those married people who had never divorced. We also saw that this varies by birth cohort, with considerable differences between Baby Boomers and the generations that preceded and followed them.

Here are some follow-up opportunities:

- Use GSS data to explore the differences between the natal families of those people who have been divorced and people who have married but never divorced. Use data on parents' occupational prestige (14, 15), education levels (17, 18) and socio-economic standing (20, 21) to identify the differences, if any. What do you conclude about the contribution of one's natal household to the possibility that one will have a marriage fail?
- Use GSS data to explore the income, occupational, and socio-economic differences between people who have been divorced and remarried and those married people who have never been divorced. What does this tell you about the effect of divorce without remarriage on one's economic standing?
(Note that this is different from Chapter 9's exercise. There we compared all people who had been divorced to people who stayed married. The suggestion here is to use a control variable to run the same comparisons with just at those divorced people who have remarried, on the one hand, and those who have not remarried, on the other.)
- Explore the same factors, plus religion, age at the birth of the first child, and any other items you think might be important, this time controlling for race.
 - Which of the patterns that we see among Whites do we also see among African Americans? Which patterns are different?
 - Is there a difference between native born interviewees and immigrants? (Use variable 38: Born USA?) How about the children of immigrants? (Use variable 39, which measures the number of the respondents' grandparents who were born outside the U.S.)
- Use the States dataset to explore the social climate in states where there is a lot of divorce. What accompanies divorce? Has this changed over the years? Take a look at poverty, inequality, education levels, religion, residential mobility, teen births, and whatever else you think might be related to divorce rates in 1980 (var. 209), 1992 (var. 134, 135), and 1998 (var. 20). Don't commit the ecological fallacy, but think through what elements of the social climate might lead divorce rates to be high or low.

4. The Sociology of Politics

The States dataset gives us a few political variables – such as the percentage of each state's voters who chose Gore or Bush in the 2000 election (variable 97), the percentage of a state's legislators who are female (variable 95), and the percentage of the population that reported voting in 1998 (variable 94)..

Here are some ways that you can put these to use:

- Chart some of factors that correlate the relative strength of a state's vote for Al Gore or George Bush. What current demographics differentiate between these states. What past demographics do so? For example: states that pay decent welfare benefits tended to vote for Al Gore; states where a high proportion of residents hunt, drive pickup trucks, and drink beer rather than wine or hard liquor tended to vote for George Bush. What other patterns can you find?
Sort out those that you think might cause people to make particular political choices from those that might merely be spurious. (For example, beer drinking did not likely lead people to vote for Bush, but wine drinkers live on the East and West Coasts, where people voted for Gore.)

- What difference does an active voting public (variable 94) make to a state's socio-political climate? What factors predict high levels of voter participation? What factors are predicted by this participation? (Use both correlation and regression analyses to figure out the differences between these causal hypotheses.)
- What difference does the presence of female politicians (variable 95) make to a state's socio-political climate? Are women better off in such states than in others? Is society healthier? Is there less poverty? Or more? Does the presence of female politicians actually cause such things, or are the correlations spurious?

The GSS dataset also records a rather large number of political attitudes, some of which we met in Chapter 8. These range from party leanings (variables 48, 49), voting behavior (50-52), attitudes toward government spending and policies (54-71), and civil liberties (72-90), to attitudes toward various current political controversies (e.g.: abortion: 167-173; policies to reverse economic inequality 220-228; "green" policies: 230-234).

Try these exercises:

- What kinds of people align with the Democratic and Republican parties? What are the various influences of race, class, gender, religion, etc on party affiliation?
- What other attitudes do such people hold? What constellations of ideas seem to go together in the minds of various party adherents?
- Which of the various political attitudes in our dataset don't divide on party lines? What, other than party, distinguishes one side from another?
- Explore some of the connections between political variables and measures of personal morality. For example, GSS variables 167-173 measure attitudes toward abortion, variables 175 and 177-182 ask about sexual attitudes, and variables 238-245 ask questions about personal sexual experiences. Do these vary along political party lines? According to liberal or conservative political views? If connections exist, are they real or are they spurious – caused by underlying demographic factors?

5. Race in American Society

Both the GSS and the States datasets have variables for race. Chapters 6 and 8 showed us how to use GSS variable 6 to measure various differences between African Americans and Whites. GSS variable 61 measures attitudes toward government spending on African Americans, variables 75-77 measure how people balance free speech against public advocacy of racism, variable 106 measures attitudes toward racial intermarriage, variables 200-206 report attitudes towards racial discrimination, and so on.

On the States side, variables 21-28 give the racial/ethnic makeup of the various states in 2000, and variables 144-148 do the same for 1990. Variables 244, 246, 256, 277, 278, 288, and 289 do so for previous years. Variables 50-56 give teen birth and infant mortality figures for Whites, Blacks, and Hispanics. Several other variables give data on immigrants, though – due to illegal immigration – these data are not as reliable as one might wish.

Here are some ideas for you:

- What attitudinal differences do you find between African Americans and Euro-Americans? How does each line up politically, socially, religiously, morally, and so on? Are there underlying patterns to people's views on the various questions listed above? Are there patterns to their views on questions of police brutality, confidence in social institutions, etc?

- Pay particular attention to views of religion (GSS variables 96-105) and personal morality (variables 167-182, 238-245), as these are often overlooked areas of racial difference.
- Which of these patterns are due to race, and which are due to other demographic factors? Explore the relative influence of race and class, race and education, race and religion, and so on? Where is race the most important line of division? Where does it merely reflect other divisions in society?
- What kinds of things happen in states with ethnically diverse populations? What kinds of things happen where the population is less diverse? Try to make sense of these differences, seeing which ones are real relationships and which are spuriously caused by other factors.
- What is the relationship between race and poverty? Use both datasets to explore this question, looking at such things as:
 - What differentiates high-poverty states from low-poverty states? Is there more unemployment? More female-headed families? Less (or more) education? More (or less) racial diversity? More divorce? Etc. Which of these factors are real and which are spurious?
 - What explains the significant difference in family income between African Americans and Euro-Americans? How much of it is caused by differences in education, occupational prestige, class standing, and so on? What other factors need to be considered?

6. The Relevance of Religion

The States dataset has only five religious variables: numbers 189-193 record (respectively) the percentage of each state's population that was irreligious, is Catholic, Baptist, Jewish, and Christian in the early 1990s. (These are not all mutually exclusive categories.) Besides our studies of the relationship between religion and suicide (Chapter 5), Catholicism and abortion (Chapter 4), and Baptists and murder (Chapter 4), this dataset lends itself to questions like the following:

- What social factors distinguished highly religious states from non-religious ones in the early 90s? Explore economics, voting, marital behavior, health and disease, poverty, employment, and so on to see what happens in states with both high and low levels of religious participation.
- Which of these correlations show actual connections with religion, either with religion in general or with one of another religious group? Which are spuriously caused by other factors?

The GSS dataset has several religious variables. Variables 96-100 and 102 cover such things as the religion with which people identify (if any), how liberal, fundamentalist, or charismatic* that religion is, how often people attend church, how religious they think they are, how often they pray. Variables 101 and 103-105 ask about specific religious beliefs. These variables let you explore the following topics:

- Is there a connection between religion and sexism? If so, what is it? Is it real, or is it the spurious result of other factors?

* The GSS uses the term "charismatic" to include both Pentecostal Protestants and Catholic Charismatics. As you can see by cross-tabulating variable 96 with variable 98, some of those who identify themselves as Jews, "others", and "nones" also claim affiliation with a charismatic movement.

- What is the connection between religious liberalism and political liberalism? Between religious and political conservatism? Can you find any surprises on this score?
- How do religious fundamentalists, moderates, and liberals differ on moral views (GSS variables 167-182, 238-245)?
- How can you explain the finding that even non-religious Americans frequently pray? (22% of those with “no religious preference” pray at least once a day.) Explore this using both the other religious variables and selected non-religious variables? What deeper cultural pattern does this show?
- On the whole, African Americans are much more religiously oriented than are Euro-Americans. They are also much more likely to have fundamentalist or charismatic (Pentecostal) views.
 - To what extent does this explain the Black/White differences in moral views?
 - What attitudes (moral or other) differentiate fundamentalist Blacks from fundamentalist Whites? (Use variable 6 as the independent variable, the attitude that you are testing as the dependent variable, and variable 97 as the control variable.)

7. Gender Differences

Chapter 7 showed us that women and men are equally sexist. Chapter 6 showed that they are equally happy. Chapter 9 told us that they do not generally vote the same way, nor do they uniformly share political attitudes. What other gender differences are there in American life?

- Women’s income is lower than men’s. Is this true even after controlling for the fact that women’s levels of education and job status are also lower? Is this true of all social classes? Of all marital statuses?
- GSS variables 208-210 report people’s attitudes toward gender discrimination. Do men and women have the same attitudes on such matters? What happens when you control for sexism (variable 166)? When you control for other factors?
- GSS variables 148-151 report how often people spend social evenings with relatives, with neighbors, with friends, and in bars. Do men and women have the same patterns of socializing? What happens when you control for marital status (variable 22)? When you control for religiosity (variable 100)? For other factors? What does this tell you about the gender roles in American society?
- Explore men’s and women’s personal moral attitudes (variables 167-182). What differences do you find? What similarities? How does marital status affect these findings? Race? Religion? What conclusions can you draw about the relative influence of each of these factors on morality?
- Run the same analysis looking at sexual behavior (variables 238-245). What conclusions can you draw?
- GSS variables 121-137 report people’s confidence in major social institutions -- religion, government, the military, banks, and so on. Are there any differences between men and women? If so, do can they be explained by men’s and women’s different levels of religiosity, education, political views, etc? When we take these variables into account, is gender still a factor in such matters, or is its relationship spurious?

8. Southern Culture

In Chapter 4, we saw that the murder rate is higher in the Old South than in other regions of the country. We also saw that this was not because of the large number of Baptists who live in that region; the connection between the percentage of a states' population that is Baptist and its murder rate is spurious. We did not, however, delve any more deeply into this phenomenon. Is the Old South really just a violent region? Or are there other factors that create its longstanding high levels of murder, assault, and other violent crimes.

Here are some follow-up opportunities:

- Use the States data to examine the South's history of violent crime, starting with the 1920s and continuing to the present day. What constants do you find? What shifts have occurred over time? (Be sure to examine the maps as well as the shifts in the distribution of murder, assault, etc. over this time period. Include an analysis of the shifting means and standard deviations of such crimes, and see what effect the South has had on them.)
- Use the States data to locate several other variables that correlate with variable 2: SOUTH. Do you see any pattern to them? What kinds of things are more common in the South than in other parts of the country? What factors – other than Southern culture – might explain them?
- Test these hypotheses with regressions, by using the factor that you are trying to explain as the dependent variable, variable 2 as one of your independent variables, one or more sociologically appropriate variables as the other independent variables. What sense can you make of your results?
- The GSS includes variables for the respondents' current area of residence (var 8: REGION) as well as the part of the country in which she or he was raised (var 34: REGION @ 16). Using these variables, compare Southerners with those from other parts of the country. How do they differ demographically? How do they differ attitudinally? What patterns can you find?
- Use control variables, T-tests/ANOVAs, and standardized cross-tabulations to determine which of these patterns are spurious and which are real. (You will have to identify sociologically plausible variables to use as independent variables, along with area of residence.)

9. Childrearing Attitudes

The General Social Survey asks people to rank 5 character traits that they think all children should have. Respondents tell the GSS their relative sense of how important it is that children obey their parents (variable 138), be popular among their peers (variable 140), think for themselves (variable 142), work hard (variable 144), and help others (variable 146). variables 139, 141, 143, 145, and 147 tell whether respondents listed the each of these character traits in their top two.

- What kinds of people think that what kinds of character traits are important? Examine gender, race, region, education, marital status, political views, etc. to see if you can find any systematic differences.
- What other attitudes accompany these beliefs? Explore such things as sexism, political liberalism or conservatism, confidence in government and other major institutions, attitudes toward government spending, beliefs about free speech, and so on. Can you identify any patterns?

10. The Other Indexes

Chapter 7 introduced us to the Sexism Index – a measure of sexist attitudes that combines people’s answers to several variables into an easy-to-use scale. Our GSS data contains several other indexes that are worth exploring. Variables 87-90 score people according to their attitudes toward various civil liberties. Variable 108 is an index of racial attitudes. Variable 120 is an index reporting trust in other people. Variables 134-137 report people’s level of confidence in major social institutions, and variable 236 is an index of people’s willingness to make personal lifestyle sacrifices for the environment.

The States dataset also has at least two indexes. Variable 195 indicates states’ relative levels of public health. It was constructed from 17 indicators that measure such things as infant mortality rates, rates of hospital insurance, etc. Variable 144 is an index of racial and ethnic diversity. High scores indicate a more diverse population, while low scores indicate a state dominated by one racial or ethnic group.

Here’s how you might explore the social correlates of some of these indexes with our data:

- What kinds of people tend to favor free speech? Are there specific demographic factors, such as age, education, or region of the country, that predict this support? Or are the factors attitudinal, such support for political liberalism?
- What kinds of people tend to support racial discrimination? Are there specific demographic factors, such as age, education, or region of the country, that predict this support? Or are the factors mainly attitudinal? Or both?
- What kinds of people are trusting of others? What other attitudes accompany such trust?
- What kinds of people have confidence in our major social institutions? What are their patterns of confidence? How does this interact with their political and/or religious views?

Using the States dataset:

- What correlates with high and low levels of public health? Investigate such things as basic demographics, economic variables, religion, voting, and so on. Can you figure out which of these factors contribute to states’ high or low scores on this index? What do your data recommend to policy makers interested in creating a healthier society for us all?
- Run a similar analysis with the index of racial diversity. What social factors correlate with racial diversity? See if you can tease out any causal chains.

These are but a few of the possible projects that you can do with Sociological Insights[®], the States and GSS datasets, and the statistical reasoning that you have learned in this book. The States dataset has many measures of poverty, health and disease, birth and death, and so on that merit exploration. The GSS dataset measures people’s attitudes toward work, the future, immigration, and so on that we have not explored. Each of these topics can generate useful insights about American society.

The point is, you are now doing real sociology! Congratulations and welcome to the adventure.

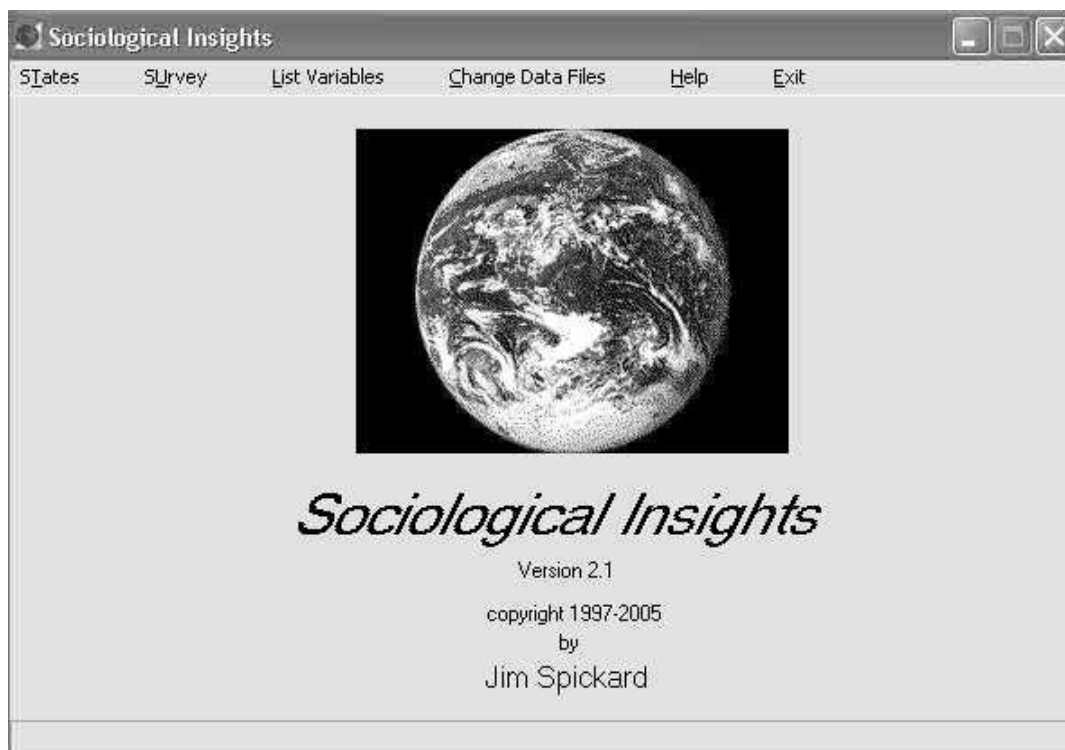
USING THE SOCIOLOGICAL INSIGHTS[®] SOFTWARE

Thinking Through Statistics is built around a special computer program, called Sociological Insights[®]. The program runs under Windows95, WindowsNT and their offspring: Windows98, Windows2000, WindowsME, WindowsXP, etc. Sorry, no Macintosh version.*

Depending on which edition of this book you have, you will either find Sociological Insights[®] on a CD attached to the inside back cover, or you will have to download it from one of the web sites listed on the back of the title page.

- If you have the CD, put the CD in your drive and open it.** You can run the program in either of two ways:
 1. You can install it on your own computer by double-clicking the “Setup” icon. This transfers the program and its data files to a “C:\Program Files\Sociological Insights\” directory on your hard drive and puts a shortcut to it in your Start Menu. Click on that shortcut to start the program.
 2. You can run it from the CD itself, by double-clicking the Sociological Insights[®] icon.
- Only the first of these options is possible if you have to download the program, though you can then copy all the files in the program directory to your own CD and run it from there.

Once you have Sociological Insights[®] running, you’ll see this opening screen:

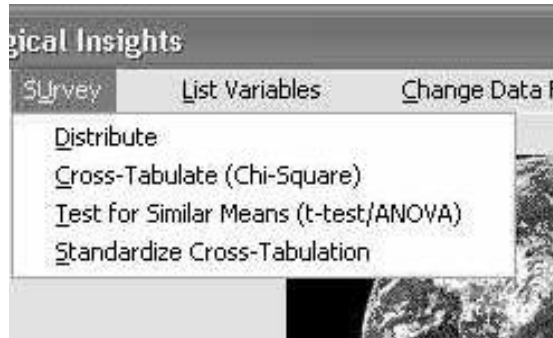


* I haven't found a Macintosh compiler that can read Delphi Pascal code. Hotshot programmers should get in touch with my publisher.

** You can double-click on “My Computer” then on the icon for your CD-ROM drive. Or you can use Windows Explorer or a similar utility. If you don't know how to do this, check your Windows manual.

Read the menu bar at the top. Use your mouse to open any menu item, or hold down the <ALT> key and type the underlined letter. <ALT>-T opens the STATES menu, <ALT>-U opens the SURVEY menu, and so on.

Here are the STATES and SURVEY menus:



The STATES menu lists all of the different things that you can do with STATES data. You can map that data across the 50 United States, which shows geographic trends (see chapter 1), and you can compare two, three, or four maps. You can create a scatterplot (chapter 3). You can correlate several variables (also chapter 3). And you can run simple regression analyses (chapters 4 & 5).

The SURVEY menu lists the things that you can do with SURVEY data. You can show a variable's distribution (chapter 6), cross-tabulate two variables (chapters 6 & 7), or control a two-variable cross-tabulation for a third variable (chapter 8). You can compare the scores that different categories of people get on a variable using a t-test or an analysis of variance, and you can control that comparison with a third variable (chapter 9). Finally, you can standardize that cross-tabulation – i.e., you can see the influence that one variable has on another, after the influence of several other variables has been removed (chapter 10).

The four other menu choices are pretty obvious. The LIST VARIABLES menu lets you list the variables for either the STATES or the SURVEY data set. The CHANGE DATA FILES menu lets you substitute a new STATES or SURVEY data file for the one with which the program starts. The HELP menu tells you how to work the program, provides copyright information and an e-mail address where you can send bug reports. And the EXIT menu item closes the program.

Each statistical routine has its own HELP menu, by the way. You are never very far from help if you need it! Each screen also has a PRINT button, so you can print what your results.

DOING STATISTICS

Each of these statistical routines is covered in one of the previous chapters, so there is no need for review here. The best way to understand what is going on is to follow the text with your computer, making sure that you understand each step before you proceed. The exercises apply a chapter's lessons to specific sociological problems.

Remember: the point of all this is not the statistics. The point is to do sociology! Statistics are merely a tool to help uncover social facts. Learning to think with statistics doesn't stop with the numbers. We use simple statistics to make thinking about social issues easier.

GOING FURTHER

Unlike most statistical software, Sociological Insights[®] is designed for teaching. It thus sacrifices flexibility for clarity. It uses real data, which you can manipulate in several ways. You can choose which variables to use. You can print its maps and charts. But you can't add data, can't recode variables, can't exclude cases, and so on. These are very useful tools, ones that you will want to use as you do more advanced quantitative research. But they muddy the water for beginning students.

Once you have understood the logic behind statistical reasoning, I suggest that you graduate from Sociological Insights[®] to a more full-featured statistical package. There are several such packages on the market, each of which has its advantages. My two favorites are MicroCase[®], now distributed by Wadsworth Publishing, and SPSS for Windows[®], published by SPSS, Inc. Each costs several hundred dollars, but they are both worth it – if you find yourself doing lots of quantitative work.

Like Sociological Insights[®], MicroCase[®] is extremely easy to use. Its menus guide one easily through its various routines, explaining options as they go. The program reads many data formats, including Excel and comma-delimited files. One can quickly exclude cases, limit ranges, and so on; recoding variables is only a bit more difficult. MicroCase[®] also publishes a large number of data sets, including all the years of the General Social Survey.

For my money, MicroCase[®] is the first choice for intermediate sociologists. (This shouldn't be surprising. MicroCase was created by sociologists and favors the kind of routines that sociologists typically use.)

That said, SPSS for Windows[®] is probably the choice for those who need raw statistical power (and have the budget to pay for it). A desktop descendant of the most famous mainframe statistical package, SPSS provides over a hundred statistical routines, plus the ability to automate data analysis through its built-in programming language. This lets you run similar analyses on different sets of data rather than wading through menus each time.

SPSS opens with a spreadsheet, which makes it easy to create new variables and recode old ones. Then one chooses one's routines from a series of menus. These are harder to navigate than are MicroCase's, but they allow more options up front. Output goes to a separate window, which one can view, print, or save.

My University licenses both Microcase[®] and the basic SPSS[®] package. The latter contains most of the routines that I need in my own research. Less frequently used routines come in add-on packages, which one must license separately – greatly increasing the cost. The only missing routine that I occasionally use is one for logistic regression – a relative of the ordinary multiple regression that we covered in chapters 4 and 5. For that, I have to use MicroCase[®]. But only SPSS[®] includes automatic stepwise regression. Having both available meets my personal needs.

Data for the 50 U.S. States

num	name	description
1	Warm Winters	Average January Low Temperature (Fahrenheit)
2	South	Degrees state capital is south of North Pole
3	Tot Pop 2000	Total population 2000 (Census)
4	% Urban 00	% of population living in metropolitan areas 2000 (Census)
5	POPchg 90-00	Percent change in population 1990-2000 (Census)
6	METchg 90-00	% change in metropolitan population 1990-2000 (Census)
7	RURchg 90-00	% change in rural population 1990-2000 (Census)
8	Med Age 2000	Median age (years) 2000 (Census)
9	% 65+ Yrs 00	% of population over 65 years old 2000 (Census)
10	Sex Ratio 00	Number of males per 100 females 2000 (Census)
11	Density 00	Population density: pop per sq. mile of land 2000 (Census)
12	# Houses 00	Total housing units 2000 (Census)
13	Vac Rate 00	Housing vacancy rate (home-owner + rental) 2000 (Census)
14	%Family 00	% of households that are families 2000 (Census)
15	%Kids 00	% of households with own children under 18 years old 2000 (Census)
16	%Married 00	Percent of households that are married couple families 2000 (Census)
17	Fem Head 00	"% of households with kids, headed by a female - no husband present 2000 (Census)"
18	Fam Size 00	Average population per family 2000 (Census)
19	Marry Rt 98	Marriage rate per 1000 population 1998 (Census)--without Nevada (79.5)
20	Divorce 98	Divorce rate per 1000 population 1998 (Census)
21	% White 00	% of population that is white 2000 including Hispanic (Census)
22	% White NL 00	% of population that is white (not including Latino) 2000 (Census)
23	% Hispanic 00	% of population that is Hispanic (any race) 2000 (Census)
24	% Black 00	% of population that is black 2000 (Census)
25	% Asian 00	% of population that is Asian 2000 (Census)--excluding Hawaii
26	% Nativ 2000	% of population that is Native American (inc. Eskimo & Aleut) -- 2000 (Newsweek)
27	%2-Race 00	% of population that is biracial 2000 (Census)
28	New Immig 98	Immigrants admitted 1998 (Census)
29	Inc/Cap 00	Personal Income per Capital 2000 (Census)
30	Dist Inc 00	% of nation's total personal income earned by state residents 2000 (Census)
31	Uneq Fam 96	Ratio of income of top 5 th of families w/ kids to income of bottom 5 th : 1996
32	% Poor 99	% of population living below the poverty level 1999 (Census)
33	% Low \$ 00	% of pop defined as low-income (under 200% of the poverty line) 2000 (Kaiser)
34	Med Inc 99	Median Income 1999 (Census)
35	Food Insec98	Food insecurity total (w/ or w/out hunger)--percent of households 1998
36	Hunger 98	Food insecurity with hunger--percent of households 1998
37	GSP 98	Gross State Product in billions of dollars 1998 (Census)
38	% Farms \$ 98	% of GSP money brought by farms - forestry - and fisheries 1998 (Census)
39	% Const \$ 98	% of GSP brought by construction 1998 (Census)
40	%Manuf \$ 98	% of GSP brought by manufacturing 1998 (Census)
41	% Trans \$ 98	% of GSP brought from transportation & public utilities 1998 (Census)
42	% Whlsl \$ 98	% of GSP brought by wholesale trade 1998 (Census)
43	%Retail \$ 98	% of GSP brought by retail trade 1998 (Census)
44	% Fin \$ 98	% of GSP brought by finance - insurance - and real estate 1998 (Census)
45	% Serv \$ 98	% of GSP brought by services 1998 (Census)
46	% Gov \$ 98	% of GSP brought by government 1998 (Census)
47	Abortion 96	Abortion rate per 1000 women 15 to 44 1996 (Census)
48	Birth Rt 99	Birth rate per 1000 population 1999 (Census)

num	name	description
49	Fertility 99	Fertility rate per 1000 women aged 15-44 years 1999 (Census)
50	Teen Birth99	Teen Births per 1000 1999 (CDC)
51	W Teen 99	White teen births per 1000 pop 1999 (CDC) (age 15-19)
52	B Teen 99	Black teen births per 1000 pop (CDC) (age 15-19)
53	H Teen 99	Hispanic teen births per 1000 pop 1999 (CDC) (age 15-19)
54	Inf Mort 99	Infant deaths per 1000 births 1999 (CDC)
55	W Inf M 99	White infant deaths per 1000 live births 1999 (CDC)
56	B Inf M 99	Black infant deaths per 1000 live births (CDC)
57	Inf Health 99	Percent of infants born with one or more of four health risks 1998 (Westat)
58	Immunize 99	% of 2-yr-olds fully immunized against preventable childhood diseases 1999 (NIS)
59	Death Rt 98	Death rate per 1000 population 1998 (Census)
60	Heart Dd 97	Deaths per 100K population due to heart disease 1997 (Census)
61	Canc Dead 97	Death rate per 100K population due to cancer 1997 (Census)
62	Car Death 97	Death rate by 100K population due to motor vehicle accidents 1997 (Census)
63	Dead Diab 97	Death rate per100K population due to diabetes mellitus 1997 (Census)
64	Dead HIV 97	Death rate per 100K population due to HIV 1997 (Census)
65	%AIDS Pop 00	Percent of the total US AIDS pop in each state 2000 (CDC)
66	AIDS Rate 00	AIDS case rate per 100K pop 2000 (Kaiser and CDC)
67	Cum AIDS	Cummulative AIDS cases thru Dec 2000 per 100K pop (Census and CDC)
68	Gun Death 99	Firearm deaths per 100K population 1999 (Kaiser Foundation)
69	Suicide 97	Death rate per 100K population due to suicide 1997 (Census)
70	Alcoholic 99	Age-adj death rate/100K pop from chronic liver disease and cirrhosis (NVSR)
71	Smoke 00	Cigarette smoking rate 2000 (CDC)
72	M Smoke 00	Cigarette Smoking Rate for males - 2000 (CDC)
73	F Smoke 00	Cigarette Smoking Rate for females - 2000 (CDC)
74	Obese 98	Overweight and obesity rate 1998
75	Mental 00	Percent reporting poor mental health in the past 30 days - 2000 (CDCP)
76	Medicaid \$98	Total Medicaid Spending per Enrollee 1998 (Kaiser)
77	% Medicare99	Medicare beneficiaries as a percent of state population 1999 (Kaiser)
78	Medicare \$00	Medicare spending per beneficiary 2000 (Kaiser)
79	Health \$ 98	Per capita personal health care expenditures 1998 (Kaiser)
80	Health/GSP97	Personal health care spending as % of the gross state product 1997 (Kaiser)
81	% Emp Ins 99	% priv sec offering health ins to employees 1999 (Kaiser)--w/out Hawaii (91%)
82	%Medicaid 00	% of the population insured by Medicaid 1999-2000 (Kaiser)
83	% No Ins 00	% of the population that is uninsured 1999-2000 (Kaiser)
84	Edu \$ 94	Educational expenditures as a % of personal income 1994 (DES)
85	HS Grad 98	Percent of 18 to 24-yr-olds with a high school credential 1998 (CPS)
86	HS Drop 98	%of students grades 9-12 not in school or secondary program 1998 (CCD)
87	College %98	% of HS grads immediately enrolled in 2-yr or 4-yr colleges 1998 (Westat)
88	HS Drink 90s	% HS students who have had 5+ drinks in a row in last 30 days 1991-99 (YRBS)
89	Disrupt 94	% of HS teachers report student disruptions interfere w/ teaching 1994 (SASS)
90	Unsafe 90s	% of students didn't go to school in past 30 days b/c felt unsafe 1993-99 (YRBS)
91	HS Guns 90s	% HS students who report carrying weapon to school last 30 dys (YRBS)
92	Teach-Pay 97	Average public school teacher salaries (in thousands) -- 1997 - Newsweek
93	Voted 1998	Percent of US citizen who report that they voted 1998 (CPS)
94	Reg Vote 98	Percent of US citizens registered to vote 1998 (CPS)
95	% Fem Leg 96	% of state legislators who are female 1996 (CWAP)
96	% Dem Leg 02	% of state legislators who belong to the Democratic Party 2002 (NCSL)

num	name	description
97	Gore/Bush 00	Number of votes for Al Gore per 1000 George Bush votes -- 2000
98	Crime 99	Crime Index 1999 (FBI--UCR)
99	Vio Crime 99	Violent crimes per 100K population (FBI--UCR)
100	Murder 99	Murder rate per 100K population 1999 (FBI--UCR)
101	Rape 99	Forcible rape rate per 100K population (FBI--UCR)
102	Robbery 99	Robbery rate per 100K population 1999 (FBI--UCR)
103	Assault 99	Aggravated assault rate per 100K population 1999 (FBI--UCR)
104	Prop Cri 99	Property Crime rate per 100K population 1999 (FBI--UCR)
105	Burglary 99	Burglary rate per 100K population 1999 (FBI--UCR)
106	Larceny 99	Larceny-theft rate per 100K population 1999 (FBI--UCR)
107	Car Theft 99	Motor vehicle theft rate per 100K population 1999 (FBI--UCR)
108	Death Row 00	Prisoners under sentence of death 12/31/00 (Bureau of Justice)
109	Executed 00	Death row inmates executed 1999 and 2000 (Bureau of Justice)
110	Crime 95	FBI Crime Index: total crimes per 100K population -- 1995
111	Vio Crime 95	FBI Crime Index: violent crimes per 100K population - 1995
112	Murder 95	FBI Crime Index: murders & non-negl. manslaughters per 100K pop 1995
113	Prop Cri 95	FBI Crime Index: property crimes per 100K population -- 1995
114	Rape 95	FBI Crime Index: forcible rapes per 100K population -- 1995
115	Robbery 95	FBI Crime Index: robberies per 100K population -- 1995
116	Assault 95	FBI Crime Index: aggravated assaults per 100K population - 1995
117	Burglary 95	FBI Crime Index: burglaries per 100K population -- 1995
118	Larceny 95	FBI Crime Index: larcenies per 100K population -- 1995
119	Car Theft 95	FBI Crime Index: motor vehicle thefts per 100K population -- 1995
120	Jail Rate 94	Number of prisoners sentences to more than 12 months per 100K pop 1994 (SCJS)
121	Total Pop 92	Total Population (in thousands) -- 1992
122	POPchg 80-90	% change in population 1980-1990 (Census)
123	% Urban 92	Percent of population living in metropolitan areas -- 1992
124	% Rural 90	Percent of population living in rural areas -- 1990
125	Birth Rt 92	Live Births per 1000 population -- 1992
126	Death Rt 92	Deaths per 1000 persons of all ages -- 1992
127	Inf Mort 92	Deaths of Infants under 1 year of age per 1000 live births -- 1992
128	Avg Life 91	Average lifetime in years 1989-91
129	Marriage Rt	Marriages per 1000 population -- 1992
130	Marriag(-NV)	Marriages per 1000 population -- MINUS NEVADA (86.1) -- 1992
131	Sex Ratio 92	Males per 100 females -- 1992
132	Stable 85-90	Percent of population age 5 & up who lived in same house 1985-90
133	% Moved 90	Percent of population age 5 & up who moved between 1985 and 1990
134	Divorce 92	Divorces per 1000 population -- 1992
135	Divrc -NV 92	Divorces per 1000 population (MINUS NEVADA: 11.1) -- 1992
136	Fem Labor 94	Participation rate of females in the labor force 1994 (SAUS)
137	GSP/CAP 94	Gross State Product per capita 1994
138	Med \$ 90-92	Three-year average of median household income -- 1990-92
139	% UNEMP 93	Average percent unemployed during 1993
140	% Foreign 90	Percent of population born in foreign countries -- 1990
141	% Poor 90-92	Three-year average percentage of persons below poverty line: 1990-92
142	% Poor 90	% of population living below the poverty level 1990 (Census)
143	Welfare Lvl	AFDC and food stamp benefits as a percent of poverty line -- 1993
144	Diversity 90	Index of Racial and Ethnic Diversity -- 1990
145	% Black 90	Percent of population checking African American/Black on 1990 census
146	% Hispan 90	Percent of population checking Hispanic origin on 1990 census
147	% As/Pac 90	% checking Asian/Pacific Islander on '90 census (w/o HAWAII: 61.8)
148	% Native 90	Percent of population checking Native American on 1990 census
149	Cops/10K 92	Full-time police officers per 10K population -- 1992
150	Jail Rate 93	Adult inmates per 1000 population -- 1993

num	name	description
151	Health 93	Overall Health Ranking (-19 to +22) -- 1993
152	Bad Heart 91	Heart Disease death rate per 100K population -- 1991
153	Strokes 91	Death rate from strokes per 100K population -- 1991
154	Cancer 91	Death rate from cancer per 100K population -- 1991
155	BreastCanc90	Age-adjusted death rate from breast cancer per 100K women 1988-1990
156	Prost Canc90	Age-adjusted death rate for prostate cancer per 100K males 1988-90
157	AIDS 92-93	AIDS cases per 100K population -- 1992-1993
158	TB rate 91	Cases of tuberculosis per 100K population -- 1991
159	Docs 92	Physicians per 100K population -- 1992
160	No Ins 90-92	Average % of persons not covered by health insurance: 1990-92
161	No Docs 93	% of population underserved by primary care physicians -- 1993
162	Shots 92-93	Percent of children vaccinated for measles - mumps - rubella -- 1992/3
163	Medicaid 92	Percent of population eligible for Medicaid -- 1992
164	Smokers 91	Percent of population age 18 & up who smoke regularly -- 1991
165	Overwght 91	Percent of population 18 and older who are overweight -- 1991
166	Car Death 92	Motor vehicle traffic deaths per 100K population -- 1992
167	Gun/Car Dead	Ratio of firearm deaths to motor vehicle deaths -- 1991
168	Homeless 90	Homeless count at shelters & streets per 100K population -- 1990
169	HS Drops 90	Percent of teens (ages 16-19) not in school -- 1990
170	HS Grads 91	Percent of pop. 25 yrs & older who have finished high school -- 1991
171	School \$ 92	Expenditures per pupil in public schools -- 1991-1992
172	% in Coll 92	Percent of population enrolled in higher education -- 1992
173	Jr. Coll 91	Enrollment in 2-year colleges as a % of undergrad enrollment -- 1992
174	Coll Min 92	Minority enrollment as a % of total college enrollment -- 1992
175	Coll Grad 91	% of pop. age 25+ who have completed college -- 1992
176	Poor Wmn 90	Percent of women age 18+ living in poverty -- 1990
177	Pay Gap 90	Av. women's income as a % of men's: professional occupations -- 1990
178	Fem Pol 93	Women as a percent of state legislators -- 1993
179	Teen Bir 91	Births to females age 15-19 per 1000 in age group -- 1991
180	Hung Kids 91	Percent of children who are hungry -- 1991
181	Kids in FH90	Percent of children living in female-headed households -- 1990
182	No Eng 90	Percent of kids aged 5-17 who do not speak English -- 1990
183	Unwed TBir	Percent of teen mother births that are unwed -- 1991
184	Abrt/Fem 92	Abortions per 1000 women aged 15-44 -- 1992
185	Abrt/Bir 92	Number of abortions per 1000 live births -- 1992
186	% Kids 92	Percent of population under age 18 -- 1992
187	% Old 90	Percent of population aged 65 & older -- 1990
188	% Old 95	Percent of population aged 65 and older -- 1995
189	No Relig 90	Percent of population with no religious identification -- 1990
190	% Cath 90	Percent of population saying they are Catholic -- 1990
191	% Baptist 90	Percent of population saying they are Baptist -- 1990
192	% Jewish 94	Percent of population reported Jewish -- 1994
193	% Christ 90	Percent reported as members/participants in Christian churches 1990
194	Voted 92	Percent of voting age population who voted in 1992
195	Health Soc93	Average of 17 indicators of social well-being (high is better) -- 1993
196	Commute 90	Average travel time to work (minutes) -- 1990
197	Hse Size 94	Average # of persons per household -- 1994
198	Teach \$ 95	Average teacher's salary (in \$1000) -- elementary & secondary -- 1995
199	Hazard 95	Number of hazardous waste sites on Superfund List -- 1995
200	Stu/Teach 89	Student/Teacher ratio: elementary & secondary -- 1989
201	Pickups 89	Number of pickup trucks per 1000 population -- 1989
202	NR/Nat 90	National Review subscriptions divided by Nation subs. -- 1990
203	Playboy 90	Playboy subscriptions per 100K population -- 1990
204	Cosmo 90	Cosmopolitan subscriptions per 100K population -- 1990
205	Liq/Cap 89	Per capita consumption of beer + wine + hard liquor (gallons) 1989

num	name	description
206	% Beer 89	Percent of alcohol consumption that was beer -- 1989
207	% Wine 89	Percent of alcohol consumption that was wine -- 1989
208	Abortion 80	Abortion rate per 1000 women aged 15-44 1980 (AGI)
209	Div 80 -NV	Divorce rate per 1000 population 1980 (VSUS)--w/out Nevada (17.3)
210	Male Home 80	Homes with no adult female -- per 1000 households -- 1980
211	Lost Kids 85	Missing children per 1000 population -- 1985
212	Hunters 82	Percent of population with hunting licenses -- 1982
213	% Poor 80	% of population below the poverty level 1980 (Census)
214	% Unemp 80	Unemployment rate--% of labor force out of work 1980 (GPEU)
215	% Moved 80	% of population residing in state for less than 5 years 1980 (Census)
216	GSP/CAP 80	Gross State Product per capita 1980
217	Edu \$ 80	Educational expenditures as a % of personal income 1980 (DES)
218	Jail Rate 80	Number of prisoners sentence to mmore than 12 months per 100K pop 1980 (SCJS)
219	Rape 82	Rapes per 100K population 1982 -- omit Alaska (85.4)
220	Murder 82	Murder rate per 100K population -- 1982
221	Murder 60	Murders per 100K population -- 1960
222	Murder 40	Murders per 100K population -- 1940
223	Murder 23	Murder convictions per 100K population -- 1923
224	Assault 82	Assaults per 100K population -- 1982
225	Assault 60	Assaults per 100K population -- 1960
226	Assault 40	Assaults per 100K population -- 1940
227	Burglary 82	Burglaries per 100K population -- 1982
228	Burglary 60	Burglaries per 100K population -- 1960
229	Burglary 40	Burglaries per 100K population -- 1940
230	Burglary 23	Burglary: # sent to prison in 1st six months of 1923 per 100K pop.
231	Larceny 82	Larcenies per 100K population -- 1982
232	Larceny 60	Larcenies per 100K population -- 1960
233	Larceny 40	Larcenies per 100K population -- 1940
234	Larceny 23	Larcenies: # sent to prison for larceny per 100K pop. -- 1923
235	Alcoholic 78	Deaths from cirrhosis of the liver per 100K population -- 1978
236	Alcoholic 60	Deaths from cirrhosis of the liver per 100K population -- 1960
237	Alcoholic 40	Deaths from cirrhosis of the liver per 100K population -- 1940
238	Alcoholic 23	Hospital admissions for alcoholism per 100K population -- 1923
239	Suicide 91	Suicides per 100K population -- 1991
240	Suicide -NV	Suicides per 100K population -- 1991 (w/o NV: 24.8)
241	Suicide 82	Suicides per 100K population -- 1982
242	Suicide 60	Suicides per 100K population -- 1960
243	Suicide 40	Suicides per 100K population -- 1940
244	%F Black 60	% of female population that is black 1960 (ICPSR)
245	# Women 60	Total number females 1960 (ICPSR)
246	%M Black 60	Percent of male population that is black 1960 (ICPSR)
247	# Men 60	Total number of males 1960
248	%Foreign 60	Percent of total population that is of foreign stock 1960 (ICPSR)
249	Tot Pop 60	Total population 1960 (ICPSR)
250	Same State60	Percent of the native population residing in the state of birth 1960 (ICPSR)
251	Priv Schl 60	Percent of elementary school children attending private school 1960 (ICPSR)
252	No House 60	Percent of married couples without their own house 1960 (ICPSR)
253	% Moved 60	% of population that moved in 5 years before 1960
254	Sex Ratio 60	Males per 100 females 1960 (ICPSR)
255	Tot Pop 40	Total population 1940 (ICPSR)
256	% Non-Wh 40	% of population belonging to a minority 1940 (ICPSR)
257	# Stores 40	Number of retail stores 1940 (ICPSR)
258	Emp/Store 40	Average number of employees per retail store 1940 (ICPSR)

num	name	description
259	# Whlsale 40	Number of wholesale businesses 1940 (ICPSR)
260	Emp/Whl 40	Number of employees per wholesale business
261	# Service 40	Number of service establishments 1940 (ICPSR)
262	Emp/Serv 40	Number of employees per service establishment 1940 (ICPSR)
263	% Unemp 40	Unemployment rate 1940 (ICPSR)
264	Part Unemp40	% of population that is partly unemployed 1940 (ICPSR)
265	%F NoSchl 40	% of females over age 25 who had no schooling 1940 (ICPSR)
266	%F Coll 40	% of females over age 25 with four or more years of college 1940 (ICPSR)
267	%M NoSchl 40	% of males over age 25 with no schooling 1940 (ICPSR)
268	%M Coll 40	% of males over age 25 who completed four or more years of college 1940 (ICPSR)
269	Tot Pop 20	Total population 1920 (ICPSR)
270	Density 20	Population per square mile 1920
271	#Farms 20	Total number of farms 1920 (ICPSR)
272	%Farms 20	% of total farms in each state 1920 (ICPSR)
273	%Farm Nat 20	% of farms belonging to native whites 1920 (ICPSR)
274	%Farm For 20	% of farms belonging to foreign-born whites 1920 (ICPSR)
275	%Farm NW 20	% of farms belonging to negroes and other non-whites 1920 (ICPSR)
276	%Farm W 20	% of farms belonging to native and non-native whites 1920 (ICPSR)
277	%F Black 20	% of female population that is black 1920 (ICPSR)
278	%M Black 20	% of male population that is black 1920 (ICPSR)
279	Sex Ratio 20	Number of males per 100 females 1920 (ICPSR)
280	No Read 20	% of population over age 21 who are illiterate 1920 (ICPSR)
281	F No Read 20	% of females over age 21 who are illiterate 1920 (ICPSR)
282	M No Read 20	% of males over age 21 who are illiterate 1920 (ICPSR)
283	B No Read 20	% of illiterates who are black 1920 (ICPSR)
284	Nat Men 1900	% of population made up native-born males 1900
285	Nat Wm 1900	% of population that is native-born women 1900 (Census)
286	Forgn M 1900	% of population that is made up of foreign-born males 1900 (Census)
287	Forgn W 1900	% of population made up of foreign-born women 1900 (Census)
288	% Blk M 1900	% of population made up of negro males 1900 (Census)
289	% Blk F 1900	% of population made up of negro females 1900 (Census)

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
1	Year						Year of survey
2	Age	18-30	31-44	45-64	65+		Age of respondent
3	Cohort	->1925	1926-43	1944-57	1958-71	1972+	Respondent birth cohort
4	Sex	Male	Female				Respondent sex
5	Race	White	Black	Other			Respondent race
6	Race B/W	White	Black				Respondent race (B/W)
7	Hispanic	Yes	No				Are you Hispanic/Latino/a?
8	Region	NthEast	South	Midwest	West		Region where respondent lives
9	Size Place	rural	town	sm city	suburb	big city	What kind of community do you live in?
10	Own Home?	Own	Rent				Do you (or your family) own your home? Or rent?
11	Working?	F/T	P/T	Unemp	Retired	Keep Hse	Is respondent working?
12	Self Emp?	Self	Not				Is respondent self-employed?
13	Occ Prest	<35	36-45	46-55	56-65	66+	Prestige of respondent occupation
14	Dad Prest	<35	36-45	46-55	56-65	66+	Prestige of respondent father occupation
15	Mom Prest	<35	36-45	46-55	56-65	66+	Prestige of respondent mother occupation
16	Degree	Not HS	HS Grad	J Coll	College	Grad Deg	Highest degree earned
17	Dad Degree	Not HS	HS Grad	J Coll	College	Grad Deg	Father highest degree
18	Mom Degree	Not HS	HS Grad	J Coll	College	Grad Deg	Mother highest degree
19	SEI	Low	Working	Middle	Prof/Own		Socio-Economic Index
20	Dad SEI	Low	Working	Middle	Prof/Own		Dad Socio-Economic Index
21	Mom SEI	Low	Working	Middle	Prof/Own		Mom Socio-Economic Index
22	Marital	Married	Widow	Divorce	Sep	Nev Mar	Current marital status
23	Living Sit	Mar/Tog	Part/Tog	Sep	No Part		What is your current living situation? Married - partnered - or not; living together or not?
24	Ev Divorce	Yes	No				Have you ever been divorced or separated? (Only those who have ever been married.)
25	Mate Work?	F/T	P/T	Unemp	Retired	Keep Hse	Does your spouse work?
26	Mate Prest	<35	36-45	46-55	56-65	66+	Prestige of spouse occupation
27	Mate Degree	Not HS	HS Grad	J Coll	College	Grad Deg	Spouse highest degree
28	Mate SEI	Low	Working	Middle	Prof/Own		Spouse Socio-Economic Index
29	Hswf/Wkwf	HseWife	Wk Wife	Wk Husb			Housewives - working wives - and working husbands (married & not retired)

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
30	# Kids	none	one	two-3	four-6	more	Number of respondent children
31	Age 1st Kid	< 18	18-22	23-28	29-34	35+	Respondent age when first kid was born
32	# Sibs	none	one	two-3	four-6	more	Number of respondent siblings
33	Home @16	rural	town	sm city	suburb	big city	Type of place lived in at age 16
34	Region @16	NthEast	South	Midwest	West		Region lived in at age 16
35	Moved 16	Same	Dif City	Dif Stat			Are you living in the same place where you lived at 16 - a diff city - or a diff state?
36	Family @16	Both	Dad/Step	Mom/Step	Sing Dad	Sing Mom	With whom did you live at age 16? (Other = missing data)
37	Why No Par	Death	Divorce				Why were you not living with both of your own parents?
38	Born USA?	Yes	No				Were you born in the USA?
39	Gran Born?	None	One	Two	Three	Four	How many of your grandparents were born outside the USA?
40	# in House	One	Two	Three	Four	Five+	How many people live in your household?
41	# Sm Kids	None	One	Two+			How many are under age six?
42	# Pre-teens	None	One	Two+			How many aged 6-12?
43	# Teens	None	One	Two+			How many aged 13-17?
44	# Adults	One	Two	Three	Four+		How many adults (18 & up)?
45	# Unrelated	None	One	Two+			How many household members are unrelated to you in any way?
46	# Earners	None	One	Two	Three+		How many household members earned money last year?
47	Fam Income	< \$15K	\$15-34K	\$35-59K	\$60-109K	\$110K+	How much money did your family earn last year from all sources?
48	Pol Party 1	Dem	Ind	Rep			Which political party do you favor? (LEANING grouped with party)
49	Pol Party 2	Dem	Ind	Rep			Which political party do you favor? (LEANING grouped with Independents)
50	Vote 96	Yes	No	Ineligible			Did you vote for President in 1996?
51	Vote Who 96	Clinton	Dole	Perot			For whom did you vote in 1996?
52	No Vote Who	Clinton	Dole	Perot	Not Know		If you did not vote in 1996 - for whom WOULD you have voted?

num	name	← category labels →					description
		1	2	3	4	5	
53	Pol View	Liberal	Moderate	Conserv			How would you describe your political views?
54	Space \$	Too Ltl	Just Rt	Too Much			Spending on space exploration is
55	Environ \$	Too Ltl	Just Rt	Too Much			Spending on environment is ...
56	Health \$	Too Ltl	Just Rt	Too Much			Spending on health is
57	Big City \$	Too Ltl	Just Rt	Too Much			Spending on big city problems is ...
58	Crime \$	Too Ltl	Just Rt	Too Much			Spending on fighting crime is ...
59	Drug War \$	Too Ltl	Just Rt	Too Much			Spending to combat drugs is ...
60	Educate \$	Too Ltl	Just Rt	Too Much			Spending on education is ...
61	Blacks \$	Too Ltl	Just Rt	Too Much			Spending to improve situation of African-Americans is ...
62	Defense \$	Too Ltl	Just Rt	Too Much			Spending on national defense is ...
63	For. Aid \$	Too Ltl	Just Rt	Too Much			Spending on foreign aid is ...
64	Welfare \$	Too Ltl	Just Rt	Too Much			Spending on welfare is ...
65	Road \$	Too Ltl	Just Rt	Too Much			Spending on highways & bridges is ...
66	Soc Sec \$	Too Ltl	Just Rt	Too Much			Spending on Social Security is ...
67	Transit \$	Too Ltl	Just Rt	Too Much			Spending on mass transit is ...
68	Parks \$	Too Ltl	Just Rt	Too Much			Spending on parks is ...
69	Childcare \$	Too Ltl	Just Rt	Too Much			Spending on childcare is ...
70	Equalize	Agree	Neutral	Disagree			Should the government do something to equalize incomes?
71	Tax Level	Too High	Right	Too Low			Is your personal federal income tax burden ...
72	Ath Spch	Allow	Not Allw				Would you allow an atheist to speak in your community?
73	Ath Teach	Allow	Not Allw				Would you allow an atheist to teach at the University?
74	Ath Book	Not Rem	Remove				Would you support removing an atheist book from the local library?
75	Rac Spch	Allow	Not Allw				Would you allow a racist to speak in your community?
76	Rac Teach	Allow	Not Allw				Would you allow a racist to teach at the University?
77	Rac Book	Not Rem	Remove				Would you support removing a racist book from your local library?
78	Comm Spch	Allow	Not Allw				Would you allow a communist to speak in your community?

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
79	Comm Teach	Allow	Not Allw				Would you allow a communist to teach at the University?
80	Comm Book	Not Rem	Remove				Would you support removing a communist book from your local library?
81	Milit Spch	Allow	Not Allw				Would you allow someone advocating military rule in the US to speak in your community?
82	Milit Teach	Allow	Not Allw				Would you allow someone advocating military rule in the US to teach at the University?
83	Milit Book	Not Rem	Remove				Would you support removing a book advocating military rule in the US from your local library?
84	Homo Spch	Allow	Not Allw				Would you allow an open homosexual to speak in your community?
85	Homo Teach	Allow	Not Allw				Would you allow an open homosexual to teach at the University?
86	Homo Book	Not Rem	Remove				Would you support removing an open homosexual book from the local library?
87	Spch Indx	Tolerant	Mixed	Intol			Speech index: from #72 #75 #78 #81 #84
88	Teach Indx	Tolerant	Mixed	Intol			Teach index: from #73 #76 #79 #82 #85
89	Book Indx	Tolerant	Mixed	Intol			Book index: from #74 #77 #80 #83 #86
90	Civ Lib Idx	Tolerant	Semi-Tol	Intol	V Intol		Civil Liberties index: from #87-89
91	Cap Pun	Favor	Oppose				Do you favor or oppose the death penalty for persons convicted of murder?
92	Gun Law?	Favor	Oppose				Do you favor or oppose a law which would require a police permit to own a gun?
93	Courts	Harsh	Right	Not Har			Do you think that local courts are too harsh - just right - or not harsh enough on criminals?
94	Grass?	Legalize	Do not				Should marijuana be legalized?
95	War in 10?	Yes	No				Do you expect the US to fight another world war in the next ten years?

num	name	← category labels →					description
		1	2	3	4	5	
96	Religion	Prot	Catholic	Jewish	Other	None	What is your religious preference?
97	Fund/Lib	Fund	Moderate	Liberal			Fundamentalist/Liberal denomination
98	Charisma	Yes	No				Are you affiliated with a charismatic movement?
99	Attend Ch	Never	LT 1/mo	LT 1/wk	1+/week		How often do you attend religious services?
100	How Relig?	Strong	Somewhat	Not Very	No Relig		How religious are you?
101	Afterlife?	Yes	No/Undec				Do you believe there is a life after death?
102	Pray?	Never	LT Wkly	1+/Week	1+/Day		How often do you pray?
103	Ban Sch Bibl	Approve	Disapp				Do you approve or disapprove of the Supreme Court ban on Bible reading in schools?
104	Is Bible?	Actual	Inspired	Old Book			What do you believe about the Bible?
105	Evil World?	Evil	Neutral	Good			Is the world basically evil - neutral - or basically good?
106	Intermar	Yes	No				Do you think there should be laws against racial intermarriage?
107	Black Push	Agree	Disagree				African-Americans should not push themselves in where they are not wanted.
108	Racism Inx	Racist	Not				Racism Index: from #106-107 (any YES = racist)
109	Integrated	Yes	No				Are there people of other races living around here?
110	Affirm Act	St Favor	Wk Favor	Wk Opp	St Opp		Do you support or oppose affirmative action?
111	Work Up	Agree	Neutral	Disagree			Do you think that Blacks should work their way up - without handouts - like various immigrants did?
112	Black Imp?	Improved	Same	Worse			Do you think that conditions for African-Americans are better - the same - or worse than a few years ago?
113	Happy?	Very	Pretty	Not Too			In general - are you very happy - pretty happy - or not too happy?
114	Hap Marr?	Very	Pretty	Not Too			In general - is your marriage very happy - pretty happy - or not too happy?
115	Health	Excellnt	Good	Fair	Poor		How is your general health?
116	Life	Exciting	Routine	Dull			In general - do you find life exciting - routine - or dull?

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
117	Helpful?	Helpful	Depends	Selfish			In general are people helpful or selfish?
118	Advantage?	Fair	Depends	Take Adv			In general do people try to take advantage of you or are they pretty fair?
119	Trust?	Trust	Depends	Careful			In general can most people be trusted or do you have to be careful around them?
120	Trust Indx	Trusting	Depends	Suspicious			Trust Index: from #117-119
121	Conf Bank	Lots	Some	Little			How much confidence do you have in the banking system?
122	Conf Biz	Lots	Some	Little			How much confidence do you have in big business?
123	Conf Relig	Lots	Some	Little			How much confidence do you have in organized religion?
124	Conf Educ	Lots	Some	Little			How much confidence do you have in the education system?
125	Conf Feds	Lots	Some	Little			How much confidence do you have in the federal government?
126	Conf Labor	Lots	Some	Little			How much confidence do you have in labor unions?
127	Conf Press	Lots	Some	Little			How much confidence do you have in the press?
128	Conf Medic	Lots	Some	Little			How much confidence do you have in the medical system?
129	Conf TV	Lots	Some	Little			How much confidence do you have in television?
130	Conf SupCt	Lots	Some	Little			How much confidence do you have in the Supreme Court?
131	Conf Sci	Lots	Some	Little			How much confidence do you have in science?
132	Conf Cong	Lots	Some	Little			How much confidence do you have in Congress?
133	Conf Mil	Lots	Some	Little			How much confidence do you have in the military?
134	Conf Idx5	Lots	MT Avg	Average	LT Avg	Little	Confidence in Major Institutions Index (5 levels)
135	Conf Idx3	Much	Average	Not Much			Confidence in Major Institutions Index (3 levels)
136	CI w/o Govt	Much	Average	Not Much			Confidence in Major Institutions other than Government
137	CI w/ Govt	Much	Average	Not Much			Confidence in Government
138	To Obey	First	Second	Third	Fourth	Fifth	Child Value: to obey (rank)
139	Obey Top	Yes	No				Child Value: to obey (in top 2?)

num	name	← category labels →					description
		1	2	3	4	5	
140	Well-Liked	First	Second	Third	Fourth	Fifth	Child Value: to be popular (rank)
141	Liked Top	Yes	No				Child Value: to be popular (in top 2?)
142	Think Self	First	Second	Third	Fourth	Fifth	Child Value: to think for self (rank)
143	Think Top	Yes	No				Child Value: to think for self (in top 2?)
144	Work Hard	First	Second	Third	Fourth	Fifth	Child Value: to work hard (rank)
145	Work Top	Yes	No				Child Value: to work hard (in top 2?)
146	Help Oth	First	Second	Third	Fourth	Fifth	Child Value: to help others (rank)
147	Help Top	Yes	No				Child Value: to help others (in top 2?)
148	Visit Kin	1-Sev/Wk	1-Sev/Mo	1-Sev/Yr	Never		How often do you spend a social evening with relatives?
149	Vis Neigh	1-Sev/Wk	1-Sev/Mo	1-Sev/Yr	Never		How often do you spend a social evening with neighbors?
150	Vis Friend	1-Sev/Wk	1-Sev/Mo	1-Sev/Yr	Never		How often do you spend a social evening with friends?
151	Bar Nights	1-Sev/Wk	1-Sev/Mo	1-Sev/Yr	Never		How often do you spend a social evening in a bar?
152	Lose Job	V Likely	Fairly	Not Too	Nt At All		How likely is it that you will lose your job or be laid off in the next 12 months?
153	Get Job	V Easy	Somewhat	Not Easy			How easy would it be to get a similar job with similar pay and benefits?
154	Like Job	Very Sat	Satisf	Dissat	Very Dis		Are you satisfied or dissatisfied with your job?
155	Work If \$\$	Work	Stop				If you had enough money to life comfortably for the rest of your life would you keep working?
156	Satis \$	V. Sat	Mostly	Not Sat			Are you satisfied or dissatisfied with your financial situation?
157	Change \$	Better	Same	Worse			In recent years has your financial situation gotten better stayed the same or worsened?
158	Rank \$	Far Belw	Below	Average	Above	Far Abov	How does your family income compare with other American families?
159	Ever Unemp	Yes	No				In the past ten years have you ever been unemployed for as long as a month?
160	Union	Yes	No				Do you or your spouse belong to a labor union?

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
161	Ahead?	Work	Equal	Luck			Do people get ahead by hard work or by luck?
162	Parent SOL	Much Bet	Better	Equal	Worse	Much Wor	How is your standard of living compared to your parents (at the same age)?
163	Kid SOL	Much Bet	Better	Equal	Worse	Much Wor	When your kids are the age you are now how do you think their standard of living will compare to yours?
164	Men Better	Agree	Disagree				Most men are better suited for politics than are most women.
165	Wife Home	Agree	Disagree				Men should work in the world and women should take care of the home and family.
166	Sexism Idx	Sexist	Not				Sexism Index: Agree to either #164 or #165 (or both) = sexist
167	Abort Def	Yes	No				Should abortion be legal in cases of serious birth defects?
168	Abort Want	Yes	No				Should abortion be legal if the woman is married and does not want any more kids?
169	Abort Hlth	Yes	No				Should abortion be legal if a pregnancy seriously jeopardizes the health of the mother?
170	Abort Poor	Yes	No				Should abortion be legal if the family cannot afford any more kids?
171	Abort Rape	Yes	No				Should abortion be legal if the pregnancy results from rape?
172	Abort Sing	Yes	No				Should abortion be legal if the woman is single and does not want to marry?
173	Abort Any	Yes	No				Should abortion be legal if the woman wants it for any reason?
174	Ideal #Kids	Zero-1	Two	Three	Four+	# Want	What do you think is the ideal number of kids for a family to have?
175	Teen BC	Yes	No				Should teens aged 14-16 have access to birth control without parental approval?
176	Div Easy?	Easier	As Is	Harder			Should divorce be easier to get - stay as is - or become harder to get?
177	Prem Sex	Always	Sometms	Not			Is premarital sex wrong?

num	name	← category labels →					description
		1	2	3	4	5	
178	Teen Sex	Always	Sometms	Not			Is sex wrong for early teens (14-16 yrs old)?
179	Xmar Sex	Always	Sometms	Not			Is extramarital sex wrong?
180	Homo Sex	Always	Sometms	Not			Is homosexual sex wrong?
181	Porn Law	Yes: All	Yes: <18	No			Should pornography be prohibited? For everyone? Just for children?
182	X-Movie	Yes	No				Have you seen an x-rated movie in the last year?
183	Spank	Yes	No				Is it sometimes necessary to discipline a child with a good hard spanking?
184	Euthenasia	Yes	No				Should doctors be allowed to end the life of a terminally sick person if s/he so requests?
185	Suic Sick	Yes	No				Person has right to end own life if has incurable disease?
186	Suic Shame	Yes	No				Person has right to end own life if has gone dishonored self or family?
187	Suic Wish	Yes	No				Person has right to end own life if is tired of living?
188	Cop Hit?	Yes	No				Are their circumstances when it is okay for a police officer to hit an adult male?
189	Cop Cuss	Yes	No				May an officer hit an adult male who curses at him/her?
190	Cop Murd	Yes	No				May an officer hit an adult male who is a murder suspect?
191	Cop Escp	Yes	No				May an officer hit an adult male who is trying to escape?
192	Cop Punch	Yes	No				May an officer hit an adult male who punches the officer?
193	Fear Walk	Yes	No				Is there anywhere in your neighborhood that you are afraid to walk?
194	Own Gun	Yes	No				Do you own any guns?
195	Hunt	Yes	No				Do you or your spouse hunt?
196	Newspaper?	1-Sev/Wk	1/Wk Les				How often do you read a newspaper?
197	Watch TV	0-1	2	3	4+		How many hours a day do you watch TV?
198	Moth Work	Agree	Disagree				A working mother can establish just as warm & secure a relationship with kids.

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
199	Presch Wk	Agree	Disagree				A preschool child is likely to suffer if his/her mother works.
200	W/B Disc	Yes	No				Differences between White and Black living situations are mainly due to discrimination.
201	W/B Able	Yes	No				Differences between White and Black living situations are mainly due to lesser inborn ability.
202	W/B Educ	Yes	No				Differences between White & Black living situations are mainly due to lack of educational opportunity.
203	W/B Will	Yes	No				Differences between White & Black living situations are mainly due to lack of motivation.
204	Race Work	Most Wht	Half		Most Blk		If employed: What is the racial composition of your workplace?
205	Rev Disc	V Likely	Somewhat		Not Like		What are the chances that a White will not get hired or promoted while a less qualified Black will?
206	Melt Pot	Distinct	Neutral		Blend		Is it better for America if racial/ethnic groups maintain separate cultures?
207	Imm +/-	Incr.	Same		Decr.		Should immigration be increased - stay the same - or decreased?
208	Disc Man	Likely	Not Like				What are the chances that a man will not get hired or promoted while a less qualified woman will?
209	Disc Woman	Likely	Not Like				What are the chances that a woman will not get hired or promoted while a less qualified man will?
210	Hire Women	Agree	Neutral		Disagree		Because of past discrimination employers should make special efforts to hire and promote qualified women.
211	Oth Lang	Yes	No				Can you speak another language than English?
212	Speak Well?	Well	Not Well				How well do you speak that language?
213	Use Lang	Never	< Wkly		Weekly	Daily	How often do you use that language?
214	Learned	Home	School		Other		Where did you learn this language?
215	Par Lang	Yes	No				Did one or both of your parents speak a language

<u>num</u>	<u>name</u>	<u>← category labels →</u>					<u>description</u>
		<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	
216	Gpar Lang	Yes	No				other than English at home? Did one or more of your grandparents speak a language other than English?
217	Official Eng	Favor	Oppose				Do you favor or oppose a law making English the official language of the U.S.?
218	Immig Idx	Help	Mixed	Hurt			Immigration Index: to what degree does immigration help or hurt the US?
219	Comp Use	Yes	No				Do you ever use a computer either at home at work or elsewhere?
220	Sat Democ	Yes	No				Are you satisfied with the way American democracy is working?
221	Good Life	Agree	Neutral	Disagree			Families like mine have a good chance of improving our lives today.
222	Rich Cause	Agree	Neutral	Disagree			The rich and powerful are the source of inequality today.
223	Need Inequal	Agree	Neutral	Disagree			Large differences in income help maintain American prosperity
224	Priv Ent	Agree	Neutral	Disagree			Private enterprise is the best way to solve American economic problems
225	Income Gap	Agree	Neutral	Disagree			Income differences in the U.S. are too large.
226	LDC Gap	Agree	Neutral	Disagree			Economic differences between rich countries and poor countries are too large.
227	LDC Tax	Agree	Neutral	Disagree			People in rich countries should make extra tax contributions to help people in poor countries.
228	Just Pay	Less	Right	Too Much			How does your current pay compare to a just wage for your skills and effort?
229	Books @ 16	0-2	ten-20	50-100	More		How many books were around your home when you were 16?
230	Green Grow	Agree	Neutral	Disagree			In order to protect the environment American needs economic growth.
231	Green Harm	Agree	Neutral	Disagree			Economic growth always harms the environment.
232	Green Price	Willing	Neutral	Unwill			Would you be willing to pay higher prices to help the environment?

2000 General Social Survey

num	name	← category labels →					description
		1	2	3	4	5	
233	Green Tax	Willing	Neutral	Unwill			Would you be willing to pay higher taxes to help the environment?
234	Green Cuts	Willing	Neutral	Unwill			Would you be willing to accept a cut in your standard of living to help the environment?
235	Recycle	Often	Seldom				Do you make a special effort to recycle cans plastics etc.?
236	Green Idx	V Willng	Mixed	V Unwill			Green Index: from items #232 #233 #234 #235
237	Anim Test	Agree	Neutral	Disagree			It is right to use animals for medical testing that might save lives.
238	Sex Part 12	None	One	Two	3-More		How many sex partners have you had in the last 12 months?
239	Sex Mate	Yes	No				Was one of these your mate or regular sexual partner?
240	SexOfSex12	Male	Both	Fem			What was the sex of your sex partners (last 12 months)?
241	Sex Freq12	No Sex	One-Two	1-3/Mon	1-3/Wk	More	How often did you have sex in the last 12 months?
242	Fem Sex	None	One	Two-3	Four-7	8+	How many female sex partners have you had since your 18th birthday?
243	Male Sex	None	One	Two-3	Four-7	8+	How many male sex partners have you had since your 18th birthday?
244	Condom	Yes	No				Did you use a condom the last time you had sex (vaginal oral or anal)?
245	Reg Sex	Yes	No				Was your last sex with your mate or regular sex partner?
246	Inj Drugs	Yes	No				Have you ever -- even once -- injected illegal drugs?
247	Crack	Yes	No				Have you ever -- even once -- used crack cocaine?
248	Vocab	0-4	Five-6	Seven-8	Nine-10		Number of vocabulary words person knows (a measure of education).
249	Zodiac	Fire	Earth	Air	Water		Birth (Sun) Sign -- by element