A DATA ANALYSIS PRIMER

Part 1

Introduction

Most science is predicated on the belief that the world is not just a random, chaotic mess, but rather things happen in certain ways and for certain reasons. Therefore, scientific analysis is often categorized by two major objectives—**to describe** and **to explain**. Scientists, by studying data they've collected, will document the order or patterns in the world (description) and the reasons behind such patterns (explanation). The latter draws from theoretical frameworks at our disposal. Social scientists are no different. We're interested in whether or not there are patterns or relationships among the social phenomena on which we collect information.

We have a variety of tools for collecting information or data; surveys are one major conduit for such information. Survey research translates the phenomena of interest into **variables.** A variable is anything that can vary (e.g., gender: male or female; degree: High School, Bachelor's, Master's, or Phd./Professional). Information on variables like gender from surveys can be entered into a computer program by assigning numerical values to people's answers. The data can then be analyzed for patterns.

A. Univariate Analysis: Analysis of One Variable

Frequency Distributions

In looking for patterns, researchers might begin by looking at just one variable. This is called "**univariate** analysis." One example of univariate analysis when a researcher looks at the distribution of cases respondents for one particular variable. This would be called a "**frequency distribution.**" For example, the researcher has 50 cases, and she wants to know how many men and how many women there are. That is, she wants to know the frequency of the distribution on gender. (Note that "male" and "female" were the responses that people could give, and hence are called variable's "response set"). Thus, the researcher can get a picture of how gender is distributed in the sample. While it doesn't take the researcher very far, this type of analysis is still very useful.

A concrete example using some actual data will help illustrate data analysis through frequency distributions. The data we are going to use is called the General Social Survey (GSS). Briefly, the GSS is an annual survey administered by the National Opinion Research Center, and it gathers data from individuals on many contemporary issues facing society, as well as obtaining basic demographic information.

The table below presents the distribution of people's answers to the question, "What is your highest degree?" in both raw numbers and percents.

Table 1: Self-Reported Highest Degree in 2002

	N	%
<High School	*439*	15.7
High School	*1,501*	53.6
Junior College	*206*	7.4
Bachelor	*435*	15.5
Graduate	*218*	7.8
Total	*2,799*	100

Doing Trend Analysis

Another useful way to look at frequency distributions is to examine how they change over time. This is called **trend** analysis. For instance, taking the example above, if we had surveys from different years that asked the same question of different people (What was their highest degree?), we could document whether or not there has been any change in this variable. In fact, the GSS has done just this. Here's an example of some trend data on highest degree using GSS data.

Table 2: Trends in Respondents Highest Degree: 2000, 1990 and 1980

	2000	**1990**	**1980**
<High School	15.7	20.6	30.4
High School	53.6	53.1	50.9
Junior College	7.4	5.5	3.1
Bachelor	15.5	14.4	10.8
Graduate	7.8	6.4	4.8
Total	100%	100%	100%

Interpreting Data—Putting the Numbers into Words

Statistical software generates results like those shown above. But while the computer does the grunt work, it is up to you to interpret what the numbers mean. **Being able to clearly and concisely interpret data analysis results for an audience is an important skill to develop for almost any occupation.** Here are a couple guidelines for interpreting data.

First, always provide the audience with some of the contextual detail (i.e., What are the data? Where do they come from? When was it gathered?, etc.).

Second, construct your interpretation from the "general to the specific." In other words, start out with a general statement and then move to discussing some specific numbers.

In our example here, we can look at a particular statistic that is often used in univariate analysis. This statistic is called the mode, and it represents the most frequently occurring value in a distribution. For the data presented in Tables 1 and 2, the mode would be the category with the largest percentage of respondents. Using the guidelines outlined above, an interpretation of Table 1 might go something like this:

Respondents from the 2000 General Social Survey are very uneven in terms of the highest degree earned. While respondents most often said they had a high school degree (53.6%), only 15.7% stated they had less than a high school degree, and only 7.8% achieved the highest degree (graduate).

Here's what we might say in looking at the trends in Table 2:

The percentage of GSS respondents that report having a high school degree increased somewhat from 1980 to 2000. However, there was a substantial drop in the number of respondents who had less than a high school degree, from 30.4% in 1980 to only 15.7% in 2000. Conversely, far more respondents achieved a bachelor degree in 2000 as compared to 1980 (7.8% versus 4.8%, respectively). Overall, the percentage of respondents who obtained a degree higher than a high school diploma increased from 1989 to 2000.

B. Bivariate Analysis: Looking for Relationships between Two Variables through Crosstabulation

Analysis of one variable is very useful, but often we want to explore **relationships** between two or more variables. One way we can do this is by using a technique called cross-tabulation. A "**cross-tab**" is a table that presents two or more variables together in the form of the distribution of one variable across the categories of another variable. The data can be presented as raw numbers or as percents. When we look at two variables at a time, it is called **bivariate** analysis; when we look at more than two variables it is called **multivariate analysis**. An example of a bivariate crosstab would be a table that showed the percentage of Whites at different levels of education as compared to Blacks. An example of a multivariate crosstab would be a table that showed the percentage of men and women at different levels of education, broken down by race (i.e. White men and women, compared to Black men and women).

Typically, in bivariate and multivariate analysis we conceptualize the relationship between the two variables in terms of one influencing the other. The language we use to capture such relationships is to call one variable the **independent or predictor variable**

(IV--it's doing the influencing) and the second variable the **dependent or outcome variable (DV--it's the one that is being influenced).** To create a crosstab, you first must decide on what two variables you think might be related to one another. Next, you must determine which is the independent and which is the dependent variable. Then, drawing from a conceptual framework, you must formulate and state a **hypothesis**. A **hypothesis** is an educated guess about what you think the relationship will be between the two (or more) variables. Hypotheses always state a **specific relationship** and specify the **nature or direction** of that relationship.

An example of a hypothesis using the two variables above would be: Whites are more likely to have Bachelors' degrees than Blacks. Here, race is the independent variable, and education is the dependent variable. In an actual research project, you would provide some theoretical model that specifies why you are making the argument that race makes a difference in educational outcomes.

One of the reasons that we might argue that Blacks are likely to have poorer educational outcomes as compared to Whites could be based on the fact that a disproportionate number of Blacks live in the city. We also know that schools are financed to a large extent based on property tax. Property tax tends to be lower in cities (for a whole bunch of reasons we cannot get into here). Thus, city schools—the schools many Black attend—tend to be under funded as compared to schools in the suburbs. Thus, it seems reasonable to suggest that Blacks may receive a poorer education than Whites, thus hampering them when it comes time to graduate or attend college.

So, given this (brief) background, we might hypothesize that Blacks are less likely to attend college. Well, what variables would we use to test our hypothesis? As our **dependent** variable, we would look at levels of educational attainment. There are lots of ways to measure educational attainment. Here we are going to look at the following categories: "less than a high school degree," "high school degree," "junior college," "Bachelor's" degree, and "graduate" degree. As for the **independent** variable, there are also lots of ways to measure race, but for simplicity's sake we are going to use a measure with only two categories: White and Black. The specific relationship that we will examine is the degree to which race is related to educational level. The specific **hypothesis** is that, *proportionally,* more Blacks have lower educational levels than Whites.

Note here that the measures we are using here are not "perfect" indications of the concepts we're interested in. For one thing, I've simplified the responses to the variables to make this example easier to follow. For another, there is no such thing as a "perfect" indicator. However, our measures are *reasonable* approximations, and social scientists frequently work with these types of variables.

Below are results generated from the GSS for the 2000 survey providing the raw numbers of Whites and Blacks for each level or category of education.

Table 3: Crosstabuation for Race and Educational Level

		Race		
		WHITE	BLACK	***Row Total***
Degree	LT High School	300	110	*410*
	High School	1,197	241	*1,438*
	Junior College	168	23	*191*
	Bachelor	374	39	*413*
	Graduate	186	17	*203*
	Column Total	*2,225*	*430*	***2,655***

However, raw numbers aren't very useful. Why? Because the number of people in each category is uneven, so to compare raw numbers doesn't give a good sense of what's going one. The fact that there are 2,225 Whites in this sample and only 430 Blacks makes a direct comparison of raw numbers impossible. So, what we have to do is convert the raw numbers into percents.

There are two ways to calculate percents, but only **one right way** when trying to investigate whether or not a relationship exists between two (or more) variables. Specifically, percents are calculated **within the categories of the independent variable**. What this means is that first you must identify whether the independent variable is located in the row or the column position.

In the example here, race is the independent or predictor variable, and it is located in the column. Therefore, to calculate percents, you use the totals at the end of the column as the **denominator** (bottom part of a fraction) and the number within each category of the dependent variable as the **numerator** (top part of a fraction). The computer will do the calculations, but you have to know what to tell the computer to use in its calculations. You can check your work (or the computer's work), by making sure the 100% are at the end of the columns. (If the independent variable had been located in the row, you would use the total at the end of the row as the denominator, and each value in the row as the numerator. The 100% would then be located at the ends of the rows.) So, the correct percentage crosstab would look like this:

Table 4: Percent Crosstabuation for Race and Educational Level

		Race	
		White	Black
Degree	LT High School	13.5	25.6

High School	53.8	56
Junior College	7.6	5.3
Bachelor	16.8	9.1
Graduate	8.4	4
Total	**100%**	**100%**

Interpreting the results

Again, while the computer does the number-crunching, ***you*** have to interpret what the numbers are "saying." To do this, it helps to remember the analytical question you were asking: Does race make a difference in educational attainment? To answer this question, you test the hypothesis that, proportionally, more Blacks have lower levels of education than Whites. The test of the hypothesis is the crosstab and an examination of the percentage differences. Specifically, to see if the data support the hypothesis or not, you must look at the percentage differences ***between*** **the categories of the independent variable.** What this means here is that you compare Blacks and Whites **for the same category of the dependent variable**. That is, among those who said that they have less than a high school degree, we must compare percentages of those who are Black and those who are White. Since our DV has more than 2 categories, the basic analysis would have you look at either of the "end" categories (e.g., "Less than High School" and "Graduate"). (Note: If you had only two categories of the dependent and independent, it wouldn't matter which category of the dependent you choose since there are only two. In saying something about one, you are automatically saying something about the other.)

Before going on to interpret the results, a note of caution first. It is important to understand that in survey research we **cannot talk about something causing something else**. All we can do is see if there is an **association or not**. That is, all we can do is look at whether or not a relationship exists between our independent variable, and our dependent variable. Another way to say this is that what we are really doing is **asking does knowing something about where people fall on one variable tell us anything about where they are likely to fall on the other variable**? So, for the table above, we want to know whether knowing something about people's race tells us anything about their educational level. What we are looking for is **whether or not there is a difference in the numbers we are comparing: a difference indicates a relationship between the two variables**.

So, what do our results show? At the lower end of education, we see that among Whites, 13.5% have less than a high school degree versus 25.6 % of Blacks. At the other end of education, 8.4% of Whites have a college degree whereas only 4% of Blacks do. In other words, there is a **difference** between the percents when we compare Whites to Blacks at the same level of educational attainment.

So, what are we to make of this? How do we interpret the findings? Basically, since there is a difference between the two percents it indicates that a **relationship does exist between race and education.** More specifically,

Table 4 shows a strong relationship between people's race and their educational level. Respondents who are White tend to have more education than Blacks. According to GSS respondents in 2000, about twenty-five percent of Blacks (25.6%) do not have a high school degree compared to a little bit more than one-tenth (13.5%) of Whites; conversely, 8.4% of Whites have a Graduate degree, as compared to have that among Blacks (4.0%).

So, does race make a difference? Yes. And it's pretty substantial. The difference at the upper end of educational attainment there is almost twice the difference between Whites and Blacks, and not quite twice at the lower end between the two groups.

A simple summary of how to frame an interpretation is that the researcher should answer the following questions:

- **Is there a relationship between the independent and dependent variables**? You're looking to see if there is "enough" of a difference in percentages between the groups to matter. Some guidelines for what is "enough" are in the third bullet.

- **What is the relationship**? Be careful how you state the relationship, because you may end up saying something that the results don't show. You always state the percent of people in one category the independent variable as compared to the percent of people in another category of the independent variable at the same category of the dependent variables. Using our example, "…8.4% of **Whites** have a graduate degree compared to 4% of **Blacks**."

- **How strong is the relationship**? Here, what you are looking at is the percent differential. In the example above, the **absolute percent difference** between 8.4% and 4.0% is 4.4%, which seems small; but in **relative** terms, the difference is more than 50%, a **huge** difference!

  One rule of thumb for gauging strength is as follow: If the difference is between 1-3%, then be hesitant about saying there's much of a relationship. If it's between 4-7%, then we can say there's a slight relationship. An 8-15% difference suggests a moderate relationship. And anything approaching 20% is a substantial relationship. This is only meant as a rough guideline. These aren't technical rules.

Part II – Making our Analyses More Complex: Multivariate Crosstabulation

While looking at relationships between two variables is very useful, we often want to look at relationships between more than two variables. This is called multivariate analysis. One way to do multivariate analysis is to create a multivariate crosstabulation.

Let us stick with our example of the analysis of the relationship between race and education. We saw that there was some sort of relationship between these two variables. But a good researcher has to ask him/herself whether there could be something else going on in trying to understand the dependent variable—educational attainment. Are there other things in addition to race that might influence the level of education a person has? Sure. Lots of things! For one, gender could influence educational attainment.

So, for a multivariate crosstab analysis, we can look at a third variable—here, gender. (Analyzing more than three variables requires sophisticated statistical techniques that are beyond our discussion here.) Before going on and looking at all three variables together, let's look at the bivariate relationship between gender and education.

Table 5: Percent Crosstabulation for Gender and Educational Level

		Sex	
		Male	Female
Degree	LT High School	15.9	15.5
	High School	51.1	55.6
	Junior College	7	7.7
	Bachelor	15.6	15.5
	Graduate	10.4	5.8
	Total	***100%***	***100%***

The results are pretty clear: women and men have about the same levels of education, except at the highest level, where men do better. So, there is, at best, a slight relationship between gender and educational level. But does this relative lack of a relationship between gender and educational level the same for Whites and Blacks? The following table presents information on the relationship between race and education level while considering gender at the same time.

Table 6: Percent Crosstabulation for Race and Educational Level by Gender

		White				**Black**	
		Sex				**Sex**	
		Male	Female			Male	Female
Degree	LT High School	14.3	12.8	**Degree**	LT High School	23.4	26.8
	High School	51.4	55.8		High School	55.2	56.5

Junior College	7.1	7.9
Bachelor	16.6	17
Graduate	10.7	6.5
Total	***100%***	***100%***

Junior College	3.2	6.5
Bachelor	11.7	7.6
Graduate	6.5	2.5
Total	***100%***	***100%***

To interpret these results, we need to compare each of these two tables to the original one. What you are looking for is whether or not the original pattern found between gender and education still holds when the third variable, race, is added. In Table 5, there was about the same proportion of men and women at each level of educational attainment, except for at the Graduate level (where about twice as many men had a graduate degree compared to women). So, what do we see when we look at White men and women at the lowest levels of educational attainment compared to Black men and women at the same educational level? We see that there is a difference in these two tables as compared to the original. Specifically, more White men than White women have less than a high school degree (14.3% versus 12.8%); and we also see that more Black women than Black men have less than a high school degree (23.4% versus 26.8%). The difference isn't a lot, but there is a difference. When we look at the highest level of education, we see that same pattern as in the original table: men, be they White or Black, are more likely to have a graduate degree than their female counterparts. In other words, Table 6 reveals a relationship between race, gender and educational attainment. Gender matters, but in different ways for Blacks as compared to Whites.

Obviously, there is a lot more that could be said about methods of analysis and various types of statistics. But that's why we offer classes in statistics and methods! But hopefully this summary will help those of you who haven't taken these classes to understand some basic terminology; and help some of you who have taken these classes to remember what you have learned.