I. Introduction to Quantitative Methods

When students in an undergraduate political science course learn that quantitative methods will be employed their first reaction often is negative. Many have had unhappy experiences in high school or college math courses, and some may have taken a statistics course that left them bored, bewildered or both. This introduction to quantitative methods is designed for the student with no background in these methods, computing, or statistics. Naturally we hope that any fears or doubts you have about your abilities in these areas will be erased, and that you will pursue more advanced knowledge. But we are committed to keeping things simple.

When we say "simple," we do not mean "simplistic" or "trivial." In fact, the techniques you will learn are fundamentally related to much more advanced and esoteric methods that people spend years mastering. Understand the simple ideas here, and you will comprehend what is essential about many of the most sophisticated techniques. So the methods we will study are extremely powerful. They are simple, however, in that you will need no quantitative or mathematical sophistication to learn them. The techniques are no more complex mathematically than the concept of a percentage. In fact, they are based on percentages and nothing more. So, if you understand what a percentage is, you have all the mathematical sophistication you need. If you know more mathematics than that, great. But you won't need those skills here.

Why learn any quantitative methods, even if they can be reduced to simple principles and techniques? We think there are two reasons for learning these methods. First, some questions of interest to political scientists cannot really be answered without them. For example, most explanations of why people vote the way they do cannot be tested without quantitative data. People take surveys, reduce the answers of their respondents to numbers, and analyze them using techniques very similar to the ones you will learn here. In fact, many of the "What?" and "Why?" questions political scientists ask, from why governments tend to create deficits, to why they go to war, to what citizens think about these things, are frequently addressed using numbers and computers. To be sure, many important (maybe even the most important) questions cannot be addressed directly using these methods. Quantitative methods cannot do everything.

The second reason we think it is important to learn quantitative methods may surprise you. We think that learning to do research with these skills is the easiest way to learn how to do research. Most students have had the experience of writing a term paper which required them to go to the library, consult sources on the problem, and write up their conclusions. Usually, this is not much fun. And, we believe, often this is not research. It is more like "pre-research" since it normally involves summarizing what other people who are authorities and who have done research think. True research involves more than that, although it may start with reviewing what others have thought and learned about a problem.

Hard Work and Good Grades: An Example

All research starts with a question. Let's start with a question close to all students' hearts, rather than worrying about political science: "Why do some students perform better in college than others?" Every student worth his meal card has an answer to that one, but we want to do research. And so far, this is not really a researchable question. To be sure, we could go to the library, talk with professors and friends, and generally collect opinions about why some students do better in school than others. But to begin our actual research, we need a theory. A theory is nothing more than a statement or set of statements about something of interest. Theories link concepts together in as explicit a way as possible. Here's a theory about why some students do better in college than others: "Hard work by students in preparation for classes and exams causes their performance in college courses to be high." This is not a fancy theory, nor is it very startling. It won't win any Nobel prizes, but it will serve to illustrate some important principles.

First, the theory states that the relationship between the two concepts, hard work and performance, is a causal one. It specifies a cause and effect relationship. Social scientists have debated for a long time whether causal theory can really be developed and tested, but for our purposes it is enough to say that developing and testing causal theory is the ultimate goal of research. The confidence we have in causal theories may be low or high, but progress in a discipline like political science depends on our ability to state, test, reject, and accept causal theories.

Second, the theory must be testable. We won't discuss all the problems theories can run into on this score. It is enough here to notice that the statement must imply some sort of comparison or variability. In this respect, our theory is not terribly well stated. For example, someone testing the theory might be tempted to go out and interview a group of hard-working students to see how well they are doing in their classes. Let's say our researcher finds that the hard-working students interviewed earned an average grade of B+ in their courses. Can she conclude that hard work causes high performance? She cannot conclude anything because she has nothing to compare her hard-working subjects with (other than her own impression that B+ performance is or is not "high"). It could be that students who don't work as hard as those in the study get lower grades. Or, they may get higher grades. Or, they may get the same grades as their more compulsive colleagues.

In order to test the theory, our researcher must have variation in the measures she is using to capture the concepts in the theory. In fact, these measures are referred to as variables. She must have some variability in her measures of how hard students work so she can compare students who work hard with those who do not work hard. A better statement of the theory – one which would emphasize this need for comparability or variability – would be, "Students who work hard perform at higher levels in their courses than students who do not work hard."

To test a theory, the concepts in it must be measurable. Concepts are just abstract ideas. To gather evidence on a concept like "wealth" we must be able to observe it. We might use "family income" or "total assets," which are things which we can observe, to measure the concept wealth. Similarly, with a concept like "hard work," we need a measure. We might ask people, for example, how many hours per week they spend studying. Class performance might be measured by the grade earned, or the GPA in the semester in question. We can diagram the process of reducing a theory to an hypothesis which states the relationship between two variables:

Theory:	Concept X --------------------->	Concept Y
	(Hard work)	(Academic Performance)

Hypothesis:	Variable x ------------------->Variable y	
	(Number of hours spent studying per week)	(GPA)

Concepts are the building blocks of theories. Theories and concepts are not themselves observable. They are abstractions. Variables are concepts that have been measured. Hypotheses state the relationship between variables. In this example the independent variable (the presumed cause) is the number of hours spent studying per week. We hope it is a good measure of the concept hard work. The dependent variable is grade point average. The dependent variable is the effect, or the thing we want to explain. Recall that our question was why some students do better in college than others. Our theory asserts that hard work has something to do with performance. We are saying that performance in college depends on hard work. Hard work is "independent" of performance in the sense that our theory does not assert that hard work depends on performance. Causality runs from hard work to performance. If our question were "Why do some students work harder than others?" hard work would be the dependent variable, and we would cast about for a theory to explain why some students work harder than others. We might hit upon motivation as an explanation. If we measured motivation, it would be the independent variable, and our measure of hard work would become the dependent variable.

You should note that although the hypothesis states an expected causal relationship between two variables, this does not mean that the researcher believes the dependent variable is totally caused by the independent variable. Thus, another theory of college performance could be that intelligence causes performance, and there is no reason why both theories cannot be correct. In fact, a more complete theory of performance would surely want to incorporate both intelligence and hard work into its explanation, not to mention motivation, difficulty of courses, etc.

It is important to realize that the causal direction indicated by a theory – which variable is independent and which is dependent – is completely a theoretical and conceptual problem. The quantitative techniques you will learn cannot answer the question of what is causing what. Do revolutions cause discontent, or does discontent cause revolutions? One of the difficult issues in stating and testing theories is figuring out what is causing what. It is mostly a matter of plausibility, and making some reasonable assumptions. For example, the time order of the variables is important. We do not usually assume that a causal process can work backwards in time. So if the amount of discontent can be observed well before the outbreak of revolution, we would be more comfortable treating discontent as the independent variable.

The purpose of testing the hypothesis is to test a theory. Therefore, we must always be concerned about how well our variables measure the concepts in the theory. If we have good measures, we can be reasonably confident of our test. But if our measures do not capture the concept we are interested in, our test may be irrelevant to our theory. Thus, we would have to ask ourselves in our example whether the variable, "number of hours studying" is a good measure of the concept, "hard work." Perhaps when we ask people how many hours they study a week, they lie to impress us. Or, maybe they include time spent daydreaming when we want to know about study time. If we have bad measures of our concepts, whatever we do to test the hypothesis won't mean much.

The Best of Times, the Worst of Times: Another Example

We are interested in how people in America are getting along economically. Specifically, we want to know whether people feel their personal financial situation is improving or slipping. We might be interested in this because we wonder about the political consequences of people's financial situation, or we may be concerned about which groups in the population are gaining or losing ground economically. One way of studying the problem would be to take a survey. We might consider surveying everyone in the national population, but such a task would be impossible. We might talk with everyone in our dormitory, or in our neighborhood, but we would quickly realize there is no guarantee that these people would be typical in any sense. Social scientists have developed techniques for sampling large populations (such as citizens of the United States of voting age) very efficiently. A random sample enables us to say something precise about the population as a whole by drawing a small sample from it (of about 1500 - 2000 cases). Drawing a truly random sample of all voting-age citizens in the U.S. is very expensive and requires a great deal of expertise. A sample is random when everyone in the population has an equal chance of being included. With a random sample of this sort, we can make statements about the population within a fairly narrow (and known) range of error.

Happily for us, a random sample of the national electorate was taken in the period immediately before and after the 1996 election. The interviews administered to the respondents in the sample were designed to solicit information which would help political scientists study how and why citizens voted, as well as what they thought about the issues and candidates in the election campaign. The examples that follow draw upon this survey.

One of the questions posed in the 1996 survey was about each respondent's financial condition.[1] The concept of interest to us (we haven't stated any kind of theory yet) is whether people's financial situation has improved or worsened. The variable we employ to measure this concept is a direct question asked of each of the respondents in the survey. Because people provide different answers to the question, we have a variable, or a measure that varies in the value it takes across the subjects in the study. One way of examining how much it varies is by looking at a frequency distribution of the variable. The variable number in the SETUPS96 file is V056.[2]

Table 1. Frequency Distribution of V056

[1]The question was, “Would you say that you were better off or worse off than you were a year ago? Respondents could give one of three answers to the question: they could say they were better off, about the same, or worse off than they were a year ago.

[2]See Charles Prysby and Carmine Scavo, *Voting Behavior: The 1996 Election* Washinton DC, The American Political Science Association, 1997. The table is constructed directly from SPSS output.

V056 Better Off Than Last Year?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Better	675	44.4	44.5	44.5
	2 Same	477	31.4	31.4	76.0
	3 Worse	364	24.0	24.0	100.0
	Total	1517	99.7	100.0	
Missing	9 NA	4	.3		
Total		1521	100.0		

A frequency distribution shows how many cases fall into each category of the variable. The categories of the variable are typically assigned numbers as well as labels to help interpret what the numbers mean. In this case, V056 is an ordinal variable which simply means the categories describe an order from less to more (excluding missing data). Respondents who said they are better off were assigned a 1. They are "less worse off" than those who said things were the same (they were assigned a 2), who were lower on the scale than those who said they were worse off who were given a 3.

The frequency distribution of V056 shows that 675 respondents classified themselves as better off than they had been a year before, whereas 364 thought they were worse off than a year before. If you add up the cases, you will find that a total of 1521 people were interviewed in this survey. Of these, 1517 answered the question, while 4 are classified as "missing."[3]

Notice that we have calculated relative frequencies or percentages. By looking at the actual frequencies, we can tell how many of our respondents were better off than a year ago. We can also tell by scanning the other categories that more were better off than were either the same or worse off. But we are not interested in the sample we have drawn for its own sake. Our interest in these 1521 people results from the fact that they comprise a random sample which can tell us about all voting age citizens of the United States in 1996. Thus, if we know that about 44 percent of our sample was better off in 1996 than they were the year before, we can say with considerable confidence that about the same proportion of the population was better off. That is, the proportions in our sample which we can measure and observe should fairly closely match the proportions in

[3]Surveys always result in some cases being classified as missing. Several reasons account for most instances of missing data. Sometimes a question simply is not appropriate for a particular respondent. If early in the interview, a respondent says he did not vote, the interviewer would not ask later on how that person voted. On a variable measuring how people voted, this respondent would be counted as missing. Sometimes a respondent is not sure of what the answer is. Survey researchers are careful not to force a respondent to answer a question since to do so would introduce error into the analysis. Thus, if the question asks for an opinion on Central American foreign policy, the respondent without an opinion is encouraged to say so. Finally, sometimes respondents just refuse to answer a question, or a question may be omitted through interviewer error.

the population which we cannot directly observe.[4]

Calculating percentages also facilitates comparisons which are the lifeblood of any test of a theory. (Remember, we still haven't stated a theory. We might have a question, though. Our question could be, "What proportion of the American electorate is better off in 1996 than in 1992?") Let's say we are interested in knowing whether there are any differences between whites and nonwhites with respect to the variable V056. We could examine the differences by comparing two frequency distributions, one for whites, and one for nonwhites:

Table 2a. Frequency Distribution for V056, Whites Only

V056 Better Off Than Last Year?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Better	522	43.5	43.5	43.5
	2 Same	389	32.4	32.4	75.9
	3 Worse	289	24.0	24.1	100.0
	Total	1199	99.9	100.0	
Missing	9 NA	1	.1		
Total		1201	100.0		

Table 2b. Frequency Distribution for V056, Nonwhites Only

V056 Better Off Than Last Year?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Better	153	48.5	48.9	48.9
	2 Same	87	27.5	27.8	76.7
	3 Worse	73	23.1	23.3	100.0
	Total	313	99.1	100.0	
Missing	9 NA	3	.9		
Total		316	100.0		

You can see that calculating percentages makes the comparisons between whites and nonwhites much easier than would be the case comparing simple frequencies. This is true because there are many more whites (1201) in the sample than nonwhites (316). So the fact that 522 whites said they were better off is hard to compare with the finding that 153 nonwhites said they were better off. But the percentages are easy to compare. When we look at the relative frequencies, we see immediately that whites were somewhat less likely to think they are better off than nonwhites. Whereas almost one half of nonwhites saw their financial situation as improved, 43.5% of whites

[4]It is possible to be a good deal more precise about the probable proportion in the population, based on the proportion in the sample. We do not address this topic – inference about a population from a random sample – in this discussion.

were better off.

Financial Condition and the Vote: How to Test a Theory

If we're going to do research, we need a theory. And if we're going to have a theory, we need a question. One of the things elections are designed to do is hold leaders accountable for disasters big and small. There are lots of reasons to think elections help promote accountability, but there are also reasons to wonder. So, our question might be, "Do elections promote accountability of leaders to the public?" If we think that accountability is something that elections achieve, we might assert by way of a theory that, "people vote for incumbents when they are doing well financially and they vote for challengers when they are not doing well financially." The reasoning, briefly stated, might go something like this: In the minds of many citizens, their financial status is at least partially due to the policies of national leaders. Those in power should promote policies which benefit the economic status of citizens. If leaders' policies fail to promote prosperity, they should be turned out of office, and someone new should be given a chance to govern. If things are going well financially, the incumbent must be doing a good job, and should be rewarded with a vote of confidence. Therefore, "people vote for incumbents when they are doing well financially and they vote for challengers (against incumbents) when they are not doing well financially."

Knowing we have data on what people thought and did in the 1996 election, we might measure our independent variable with V056. We know there is variation because we have examined the frequency distribution. Thus, we know that we will have people who were doing better financially to compare with those who were doing worse. Our dependent variable will be whether people voted for the incumbent, Bill Clinton, or for one of the challenger candidates, Bob Dole or Ross Perot.[5] We can summarize our hypothesis, then, as follows:

Independent Variable		Dependent Variable
x		y
V056		DIVOTE
Respondent's Financial Condition	--------------------------->	Voting Choice in 1996

That is, we believe that people's financial condition causes their vote. We use survey questions that ask respondents about their financial condition compared with the year before the election, and how they voted in the 1996 presidential election.[6]

How do we test the hypothesis? If we are going to accept the hypothesis, we must find a

---

[5]We use a recoded version of V002, Presidential Vote, which combines voted for Dole or for Perot into a single category. We call this new variable, DIVOTE (for "dichotomized vote").

[6]The choice of measures is not trivial. We could explore the effects of financial condition on voting for other offices such as the U.S. House of Representatives, on the possibility that voters are more likely to hold presidents accountable for their financial condition than House members, or we could examine the effects of respondents judgments about the state of the economy, rather than their own personal financial conditions. These and other questions have been examined in considerable depth in the literature on this question.

statistical relationship between our independent and dependent variables. If we find a statistical relationship, the case for our hypothesis is strengthened (although not proven -- see Part II). If we find no relationship between the two variables, we would reject the hypothesis. Remember, the goal is to see if the evidence supports our theory. We must measure variables which we think are related to our concepts. If the evidence is consistent with our theory (if there is a statistical relationship between our variables), we have greater confidence in our theory, and we have a partial answer to the question about whether elections promote accountability.[7]

What is a statistical relationship? A statistical relationship occurs when two variables co-vary -- when there is a difference in one variable that is associated with variation on the other. In this example, our hypothesis anticipates a relationship because it expects that as financial condition worsens, the tendency to vote for the challenger increases. That is, those whose financial condition is worse will be more likely to vote for a challenger than those whose financial condition is better.

We examine the data for a statistical relationship using a technique called crosstabulation. The idea is to build a table defined by both variables where each case (each respondent) is given a place in the table which is defined by the value given to her responses to the two variables that define the rows and columns of the table. Below, we present the table which reports the crosstabulation of DIVOTE with V056:

Table 3. Crosstabulation of DIVOTE and V056.

			V056 Better Off Than Last Year?			Total
			1 Better	2 Same	3 Worse	
DIVOTE	1.00	incumbent	302	182	105	589
			60.2%	49.3%	42.5%	52.7%
	2.00	challenger	200	187	142	529
			39.8%	50.7%	57.5%	47.3%
Total			502	369	247	1118
			100.0%	100.0%	100.0%	100.0%

First, notice that this crosstabulation table (often referred to as a "crosstab") is really nothing more than three frequency distributions side by side. The independent variable defines the columns of the table, and the frequency distributions of the dependent variable (DIVOTE) are presented in the rows for each category of V056. Thus, 60.2 percent of those who thought they were better off in 1996 voted for the incumbent, Bill Clinton, and 39.8 percent voted for Bob Dole or Ross Perot. These percentages are important because they permit us to compare the frequency distribution of those whose financial situation has gotten better with the DIVOTE frequency

[7]This may sound overly tentative. Partly, this is due to the fact that the direction of causality cannot be demonstrated with the data. More importantly, however, there are other possible explanations for a relationship between two variables. These explanations, some of which will be examined in Part II, are not necessarily consistent with the hypothesis that x causes y.

distributions of those whose situation is the same as or worse than the year before. The percentages are equivalent even though the numbers of cases of those whose situation is better (N=502) is different from the numbers of cases in the other two categories of V056. This, then, is exactly analogous to our example above where we compared the frequency distributions of whites with nonwhites. In fact, the results we got in that example would be identical to a crosstabulation of V056 with a dichotomized measure of race (white, nonwhite).

Comparing the percentages of those voting for Clinton across the three categories of V056, we see immediately that there is indeed a statistical relationship between our two variables. Those whose financial situation was better were most likely to vote for the incumbent, Clinton. Those whose situation was the same were less likely to vote for Clinton, and those whose finances had suffered were least likely to vote for Clinton. This is just as the hypothesis expects. Thus we can conclude the evidence assembled in this crosstabulation supports a theoretical expectation consistent with the idea that elections serve to promote accountability of leaders to the electorate, with respect to voters' personal financial affairs.

A crosstabulation is sometimes referred to as a joint frequency distribution. The cells in the table are composed of the numbers of cases which meet the joint condition implied by the categories of both variables. Thus, 302 respondents in the 1996 survey were better off financially and voted for Clinton; 200 of those interviewed were better off and voted for Dole or Perot, and so on through the 6 cells defined by the table.

You might ask why the percentages reported in the table are based on the totals in the column (e.g., 60.2% = 302/502). There are two other ways percentages could be calculated. Percentages could be calculated on the row totals (e.g. 302/589 = 51.3%). This would tell us that 51.3 percent of those who voted for Clinton were better off. In this case, the relative frequency distribution would be based upon the totals of each category of the dependent variable. The other way of calculating the cell percentages would be to ask what percentage of the total were better off and voted for Clinton (302/1118 = 27.0%). In this way of calculating the percentages, the cell entries become the percentage of the total sample meeting both conditions defined by the table. The reason we must calculate the percentages within each category of the independent variable is because we are interested in comparing frequency distributions of the dependent variable (What percentage voted for Clinton?) as economic condition varies. That comparison is possible only when we make the frequency distributions of the dependent variable comparable for each value of the independent variable. We do this by calculating the percentages for each category of the dependent variable within the categories of the independent variable.[8]

So far, we have just asked whether there is or is not a statistical relationship between our variables. But the strength of a statistical relationship can vary from weak to strong. Many times it is useful to compare the relationship we find in one case with others. For example, we may want to know whether there is a stronger relationship between financial condition and vote than

[8]Setting up the table so that the percentages are correctly computed is important to do when you are running the software that generates the crosstab.

between opinion on Central America and vote. Or, we may want to know, against some generally accepted standard, just how strong the relationship we have found is. One way of dealing with this is to ask two questions: "What does a perfect relationship look like?" and, "What does the complete absence of a relationship look like?" Hypothetical data can illustrate the answers. A perfect relationship between DIVOTE and V056 would exist when voting choice is completely dependent on financial condition (we reduce the coding on V056 to a dichotomy to simplify things):

Table 4. An Example of a Perfect Relationship (Hypothetical)

	V056	1. Better	2. Worse
DIVOTE	1. Incumbent	100%	0%
	2. Challenger	0%	100%
		100%	100%

In table 4, all of those whose financial condition was better voted for the incumbent, and every single one of those whose financial condition worsened voted for the challenger. Knowing financial condition permits perfect prediction about how a respondent voted.

The complete absence of a relationship indicates no predictability or covariation whatsoever:

Table 5. An Example of the Absence of a Relationship (Hypothetical)

	V056	1. Better	2. Worse
DIVOTE	1. Incumbent	60%	60%
	2. Challenger	40%	40%
		100%	100%

There is no relationship in this table because there is no difference in the frequency distributions of DIVOTE between those whose financial situation is better and those whose financial situation is worse. In general, the greater the differences in the frequency distributions, the stronger the relationship. Thus, in the example of the perfect relationship, the differences in voting between the two categories of V056 are as great as they can be (100%). With no relationship, there are no differences (0%).

A summary statistic which describes the strength of the relationship between two variables is the correlation coefficient. There are many different coefficients, each with slightly different properties, but we treat them only very generally. Our examples in the next section employ Kendall's Tau. Tau is a useful summary of a relationship, but it can never substitute for examining the crosstabulation carefully and studying the percentages you find there.

The Tau correlation measures the strength of a relationship between two variables. Like all correlation coefficients, it varies in absolute value between 0 and 1.0. In our examples above, the perfect relationship would yield a correlation of 1.0, while the absence of a relationship would result in a correlation of 0.0. In working with survey data, even "weak" correlations of .10 - .20 can be interesting, and the strongest correlations seldom exceed the .60 - .70 range. Frequently, one is most interested not in the absolute value of the correlation (is the relationship strong or weak?), but the relative value compared to some other relationship. Calculating the correlation coefficient makes sense only if the variables being analyzed can be considered ordinal in their coding. Thus, it must be possible to think of the numbers associated with each category as capturing "more" or "less" of some quantity. If the numbers do not have that meaning -- if there is no order implied by the coding of the variables -- it is not appropriate to calculate a correlation coefficient.

Summary

By now you should have a pretty good idea of the basic steps involved in quantitative research. You start with a question such as, "Why do people vote the way they do?" After thinking and reading about the question, you come up with a theory. A theory states a causal relationship between (or among) concepts. A good theory will be reasonably precise about what each concept means, and why the relationships that are specified are expected. A hypothesis states the expected relationship between variables. Variables are concepts that have been measured. The independent variable measures the presumed cause, the dependent variable measures the presumed effect. The hypothesis expects a statistical relationship, or correlation, between the two variables. The stronger the relationship, the greater the difference in the frequency distributions of the dependent variable compared across categories of the independent variable.

Key Concepts

Theory
Variable
Hypothesis
Independent variable
Dependent variable
Random sample
Frequency distribution
Missing data
Relative frequency
Ordinal variable
Statistical relationship
Covariation
Crosstabulation
Correlation coefficient