Public Opinion on Social Issues -- 1975-2010

Elizabeth N. Nelson and Edward E. Nelson, California State University, Fresno

Chapter 2
Survey Research Design and Quantitative Methods of Analysis for Cross-Sectional Data

Almost everyone has experience with surveys. Market surveys ask respondents whether they recognize products and their feelings about them. Political polls ask questions about candidates for political office or opinions related to political and social issues. Needs assessments use surveys that identify the needs of groups. Evaluations often use surveys to assess the extent to which programs achieve their goals.

Survey research is a method of collecting information by asking questions. Sometimes interviews are done face-to-face with people at home, in school, or at work. Other times questions are sent in the mail for people to answer and mail back. Increasingly, surveys are conducted by telephone and over the internet.

SAMPLE SURVEYS

Although we want to have information on all people, it is usually too expensive and time consuming to question everyone. So we select only some of these individuals and question them. It is important to select these people in ways that make it likely that they represent the larger group.

The **population** is all the objects in which we are interested. Often populations consist of individuals. For example, a population might consist of all adults living in California. But it may also be geographical areas such as all cities with populations of 100,000 or more. Or we may be interested in all households in a particular area. A **sample** is the subset of the population involved in a study. In other words, a sample is part of the population. The process of selecting the sample is called **sampling**. The idea of sampling is to select part of the population to represent the entire population.

The United States Census and the American Community Survey are good examples of sampling. The census tries to enumerate all residents every ten years. In 2000, approximately one out of every six households was given a longer questionnaire and five out of every six households were given a shorter version . Information from this sample was used to make inferences about the population. In 2010 evryone received the short form (about 10 questions) in the census. The American Community survey is conducted every year and uses sampling to provide up to date information about communities.

Political polls also use samples. To find out how potential voters feel about a particular political race, pollsters select a sample of potential voters. This module uses opinions from a sample of adults (18+) living in the United States collected at several points in time.

Since a survey can be no better than the quality of the sample, it is essential to understand the basic principles of sampling. There are two types of sampling-probability and nonprobability. A **probability sample** is one in which each individual in the population has a known, nonzero, chance of being selected in the sample. The most basic type is the **simple random sample**. In a simple random sample, every individual (and every combination of individuals) has the same chance of being selected in the sample. This is the equivalent of writing each person's name on a piece of paper, putting them in plastic balls, putting all the balls in a big bowl, mixing the balls thoroughly, and selecting some predetermined number of balls from the bowl. This would produce a simple random sample.

The simple random sample assumes that we can list all the individuals in the population, but often this is impossible. If our population were all the households or residents of California, there would be no list of the households or residents available, and it would be very expensive and time consuming to construct one. In this type of situation, a **multistage cluster sample** would be used. The idea is very simple. If we wanted to draw a sample of all residents of California, we might start by dividing California into large geographical areas such as counties and selecting a sample of these counties. Our sample of counties could then be divided into smaller geographical areas such as blocks and a sample of blocks would be selected. We could then construct a list of all households for only those blocks in the sample. Finally, we would go to these households and randomly select one member of each household for our sample. Once the household and the member of that household have been selected, substitution would not be allowed. This often means that we must call back many times, but this is the price we must pay for a good sample.

Telephone samples often use a technique called **random-digit dialing**. With random-digit dialing, phone numbers are dialed randomly within working exchanges. Numbers are selected in such a way that all areas have the proper proportional chance of being selected in the sample. Random-digit dialing makes it possible to include numbers that are not listed in the telephone directory and households that have moved into an area so recently that they are not included in the current telephone directory. More and more households do not have or use a landline but rely exclusively on their cell phone. This has greatly complicated telephone sampling.

A **nonprobability sample** is one in which each individual in the population does not have a known chance of selection in the sample. There are several types of nonprobability samples. For example, magazines often include questionnaires for readers to fill out and return. Talk shows and the media often use "instant polls" in which people are asked to call in or answer a short survey on the web to get information about attitudes and opinions. These are **volunteer samples** since respondents self-select themselves into the sample (i.e., they volunteer to be in the sample). Another type of nonprobability sample is a **quota sample**. Survey researchers may assign quotas to interviewers. For example, interviewers might be told that half of their respondents must be female and the other half male. This is a quota on sex. We could also have quotas on several variables (e.g., sex and race) simultaneously.

Probability samples are preferable to nonprobability samples. First, they avoid the dangers of what survey researchers call "systematic selection biases" which are inherent in nonprobability samples. For example, in a volunteer sample, particular types of persons might be more likely to volunteer. Perhaps highly-educated individuals are more likely to volunteer to be in the sample and this would produce a systematic selection bias in favor of the highly educated. In a probability sample, the selection of the actual cases in the sample is left to chance. Second, in a probability sample we are able to estimate the amount of sampling error (our next concept to discuss).

We would like our sample to give us a perfectly accurate picture of the population. However, this is unrealistic. Assume that the population is all employees of a large corporation, and we want to estimate the percent of employees in the population that is satisfied with their jobs. We select a simple random sample of 500 employees and ask the individuals in the sample how satisfied they are with their jobs. We discover that 75 percent of the employees in our sample are satisfied. Can we assume that 75 percent of the population is satisfied? That would be asking too much. Why would we expect one sample of 500 to give us a perfect representation of the population? We could take several different samples of 500 employees and the percent satisfied from each sample would vary from sample to sample. There will be a certain amount of error as a result of selecting a sample from the population. We refer to this as **sampling error**. Sampling error can be estimated in a probability sample, but not in a nonprobability sample.

It would be wrong to assume that the only reason our sample estimate is different from the true population value is because of sampling error. There are many other sources of error called **nonsampling error**. Nonsampling error would include such things as the effects of biased questions, the tendency of respondents to systematically under or overestimate individual characteristics such as age and behaviors such as voting, the exclusion of certain types of people from the sample (e.g., those without phones, those without permanent addresses, those we are never able to contact, those who refuse to answer our questions), or the tendency of some respondents to systematically agree to statements regardless of the content of the statements. In some studies, the amount of nonsampling error might be far greater than the amount of sampling error. Notice that sampling error is random in nature, while nonsampling error may be nonrandom producing systematic biases. We can estimate the amount of sampling error (assuming probability sampling), but it is much more difficult to estimate nonsampling error. We can never eliminate sampling error entirely, and it is unrealistic to expect that we could ever eliminate nonsampling error. It is good research practice to be diligent in seeking out sources of nonsampling error and trying to minimize them.

DATA ANALYSIS: Examining Variables One at a Time (Univariate Analysis)

The rest of this chapter will deal with the analysis of survey **data**. Data analysis involves looking at variables or "things" that vary or change. A **variable** is a characteristic of the individual (assuming we are studying individuals). The answer to each question on the survey forms a variable. For example, sex is a variable-some individuals in the sample are male and some are female. Age is a variable; individuals vary in their ages.

Looking at variables one at a time is called **univariate analysis**. This is the usual starting point in analyzing survey data. There are several reasons to look at variables one at a time. First, we want to describe the data. How many of our sample are men and how many are women? How many are

African-Americans and how many are white? What is the distribution by age? How many say they are going to vote for Candidate A and how many for Candidate B? How many respondents agree and how many disagree with a statement describing a particular opinion?

Another reason we might want to look at variables one at a time involves recoding. **Recoding** is the process of combining categories within a variable. Consider age, for example. In the data set used in this module, age varies from 18 to 89 (i.e., 89 is used for everyone 89 or older), but we would want to use fewer categories in our analysis, so we might combine age into age 18 to 29, 30 to 49, and 50 and over. We might want to combine African Americans with the other races to classify race into only two categories-white and nonwhite. Recoding is used to reduce the number of categories in the variable (e.g., age) or to combine categories so that you can make particular types of comparisons (e.g., white versus nonwhite).

The frequency distribution is one of the basic tools for looking at variables one at a time. A **frequency distribution** is a set of categories and the number of cases in each category. **Percent distributions** show the percentage in each category. Table 2.1 shows frequency and percent distributions for two hypothetical variables-one for sex and one for willingness to vote for a woman candidate. Begin by looking at the frequency distribution for sex. There are three columns in this table. The first column specifies the categories-male and female. The second column tells us how many cases there are in each category, and the third column converts these frequencies into percents.

Table 2.1 -- Frequency and Percent Distributions for Sex and Willingness to Vote for a Woman Candidate (Hypothetical Data)

Sex			**Voting Preference**			
Category	Freq.	Percent	Category	Freq.	Percent	Valid Percent
Male	380	40.0	Willing to Vote for a Woman	460	48.4	51.1
Female	570	60.0	Not Willing	440	46.3	48.9

Total	950	100.0	to Vote for a Woman			
			Refused	50	5.3	Missing
			Total	950	100.0	100.0

In this hypothetical example, there are 380 males and 570 females or 40 percent male and 60 percent female. There are a total of 950 cases. Since we know the sex for each case, there are no **missing data** (i.e., no cases where we do not know the proper category). Look at the frequency distribution for voting preference in Table 2.1. How many say they are willing to vote for a woman candidate and how many are unwilling? (Answer: 460 willing and 440 not willing) How many refused to answer the question? (Answer: 50) What percent say they are willing to vote for a woman, what percent are not, and what percent refused to answer? (Answer: 48.4 percent willing to vote for a woman, 46.3 percent not willing, and 5.3 percent refused to tell us.) The 50 respondents who didn't want to answer the question are called missing data because we don't know which category into which to place them, so we create a new category (i.e., refused) for them. Since we don't know where they should go, we would want a percentage distribution considering only the 900 respondents who answered the question. We can determine this easily by taking the 50 cases with missing information out of the base (i.e., the denominator of the fraction) and recomputing the percentages. The fourth column in the frequency distribution (labeled "valid percent") gives us this information. Approximately 51 percent of those who answered the question were willing to vote for a woman and approximately 49 percent were not.

With these data we will use frequency distributions to describe variables one at a time. There are other ways to describe single variables. The mean, median, and mode are averages that may be used to describe the central tendency of a distribution. The range and standard deviation are measures of the amount of variability or dispersion of a distribution. (We will not be using measures of central tendency or variability in this module.)

Exploring the Relationship Between Two Variables (Bivariate Analysis)

Usually we want to do more than simply describe variables one at a time. We may want to analyze the relationship between variables. Morris Rosenberg (1968:2) suggests that there are three types of relationships: "(1) neither variable may influence one another .... (2) both variables may influence one another ... (3) one of the variables may influence the other." We will focus on the third of these types which Rosenberg calls "asymmetrical relationships." In this type of relationship, one of the variables (the **independent variable**) is assumed to be the cause and the other variable (the **dependent variable**) is assumed to be the effect. In other words, the independent variable is the variable that influences the dependent variable.

For example, researchers think that smoking causes lung cancer. The statement that specifies the relationship between two variables is called a **hypothesis** (see Hoover and Donovan, 2011, for a more extended discussion of hypotheses). In this hypothesis, the independent variable is smoking (or more precisely, the amount one smokes) and the dependent variable is lung cancer. Consider another example. Political analysts think that income influences voting decisions, that rich people vote differently from poor people. In this hypothesis, income would be the independent variable and voting would be the dependent variable.

In order to demonstrate that a **causal relationship** exists between two variables, we must meet three criteria: (1) there must be a statistical relationship between the two variables, (2) we must be able to demonstrate which one of the variables influences the other, and (3) we must be able to show that there is no other alternative explanation for the relationship. As you can imagine, it is impossible to show that there is no other alternative explanation for a relationship. For this reason, we can show that one variable does not influence another variable, but we cannot prove that it does. We can only show that a causal relationship is plausible or credible. In this chapter, we will focus on the first two criteria and leave this third criterion to the next chapter.

In the previous section we looked at the frequency distributions for sex and voting preference. All we can say from these two distributions is that the sample is 40 percent men and 60 percent women and that slightly more than half of the respondents said they would be willing to vote for a woman, and slightly less than half are not willing to. We cannot say anything about the relationship between sex and voting preference. In order to determine if men or women are more likely to be willing to vote for a woman candidate, we must move from univariate to bivariate analysis.

A **crosstabulation** (or **contingency table**) is the basic tool used to explore the relationship between two variables. Table 2.2 is the crosstabulation of sex and voting preference. In the lower right-hand corner is the total number of cases in this table (900). Notice that this is not the number of cases in the sample. There were originally 950 cases in this sample, but any case that had missing information on either or both of the two variables in the table has been excluded from the table. Be sure to check how many cases have been excluded from your table and to indicate this figure in your report. Also be sure that you understand why these cases have been excluded. The figures in the lower margin and right-hand margin of the table are called the marginal distributions. They are simply the frequency distributions for the two variables in the table. Here, there are 360 males and 540 females (the marginal distribution for the column variable-sex) and 460 people who are willing to vote for a woman candidate and 440 who are not (the marginal distribution for the row variable-voting preference). The other figures in the table are the cell frequencies. Since there are two columns and two rows in this table (sometimes called a 2 x 2 table), there are four cells. The numbers in these cells tell us how many cases fall into each combination of categories of the two variables. This sounds complicated, but it isn't. For example, 158 males are willing to vote for a woman and 302 females are willing to vote for a woman.

Table 2.2 -- Crosstabulation of Sex and Voting Preference (Frequencies)

	Sex		
Voting Preference	Male	Female	Total
Willing to Vote for a Woman	158	302	460
Not Willing to Vote for a Woman	202	238	440
Total	360	540	900

We could make comparisons rather easily if we had an equal number of women and men. Since these numbers are not equal, we must use percentages to help us make the comparisons. Since percentages convert everything to a common base of 100, the percent distribution shows us what the table would look like if there were an equal number of men and women.

Before we percentage Table 2.2, we must decide which of these two variables is the independent and which is the dependent variable. Remember that the independent variable is the variable we think might be the influencing factor. The independent variable is hypothesized to be the cause, and the dependent variable is the effect. Another way to express this is to say that the dependent variable is the one we want to explain. Since we think that sex influences willingness to vote for a woman candidate, sex would be the independent variable.

Once we have decided which is the independent variable, we are ready to percentage the table. Notice that percentages can be computed in different ways. In Table 2.3, the percentages have been computed so that they sum down to 100. These are called **column percents**. If they sum across to 100, they are called **row percents**. If the independent variable is the column variable, then we want the percents to sum down to 100 (i.e., we want the column percents). If the independent variable is the row variable, we want the percents to sum across to 100 (i.e., we want the row percents). This is a simple, but very important, rule to remember. We'll call this our **rule for computing percents**. Although we often see the independent variable as the column variable so the table sums down to 100 percent, it really doesn't matter whether the independent variable is the column or the row variable. In this module, we will put the independent variable as the column variable. Many others (but not everyone) use this convention. Please do this when you write your report.

Table 2.3 -- Voting Preference by Sex (Percents)

Voting Preference	Male	Female	Total
Willing to Vote for a Woman	43.9	55.9	51.1
Not Willing to Vote for a Woman	56.1	44.1	48.9
Total Percent	100.0	100.0	100.0
(Total Frequency)	(360)	(540)	(900)

Now we are ready to interpret this table. Interpreting a table means to explain what the table is saying about the relationship between the two variables. First, we can look at each category of the independent variable separately to describe the data and then we compare them to each other. Since the percents sum down to 100 percent, we compare across. The **rule for interpreting percents** is to compare in the direction opposite to the way the percents sum to 100. So, if the percents sum down to 100, we compare across, and if the percents sum across to 100, compare down. If the independent variable is the column variable, the percents will **always** sum down to 100. We can look at each category of the independent variable separately to describe the data and then compare them to each other-describe down and then compare across. In Table 2.3, row one shows the percent of males and the percent of females who are willing to vote for a woman candidate--43.9 percent of males are willing to vote for a woman, while 55.9 percent of the females are. This is a difference of 12 percentage points. Somewhat more females than males are willing to vote for a woman. The second row shows the percent of males and females who are not willing to vote for a woman. Since there are only two rows, the second row will be the complement (or the reverse) of the first row. It shows that males are somewhat more likely to be unwilling to vote for a woman candidate (a difference of 12 percentage points in the opposite direction).

When we observe a difference, we must also decide whether it is significant. There are two different meanings for significance-statistical significance and substantive significance. **Statistical significance** considers whether the difference is great enough that it is probably not due to chance factors. **Substantive significance** considers whether a difference is large enough to be important. With a very large sample, a very small difference is often statistically significant, but that difference may be so small that we decide it isn't substantively significant (i.e., it's so small that we decide it doesn't mean very much). We're going to focus on statistical significance, but remember that even if a difference is statistically significant, you must also decide if it is substantively significant.

Let's discuss this idea of statistical significance. If our population is all adults in the United States, we want to know if there is a relationship between sex and voting preference in this population. All we have is information about a sample from the population. We use the sample information to make an inference about the population. This is called **statistical inference**. We know that our sample is not a perfect representation of our population because of **sampling error**. Therefore, we would not expect the relationship we see in our sample to be exactly the same as the relationship in the population.

Suppose we want to know whether there is a relationship between sex and voting preference in the population. It is impossible to prove this directly, so we have to demonstrate it indirectly. We set up a hypothesis (called the **null hypothesis**) that says that sex and voting preference are not related to each other in the population. This basically says that any difference we see is likely to be the result of random variation. If the difference is large enough that it is not likely to be due to chance, we can reject this null hypothesis of only random differences. Then the hypothesis that they are related (called the **alternative or research hypothesis**) will be more credible.

Table 2.4 -- Computation of Chi Square Statistic

Column 1	Column 2	Column 3	Column 4	Column 5
f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
158	184	-26	676	3.67
202	176	26	676	3.84
302	276	26	676	2.45
238	264	-26	676	2.56
12.52 = chi square				

In the first column of Table 2.4, we have listed the four cell frequencies from the crosstabulation of sex and voting preference. We'll call these the **observed frequencies** (f_o) because they are what we observe from our table. In the second column, we have listed the frequencies we would expect if, in fact, there is no relationship between sex and voting preference in the population. These are called the **expected frequencies** (f_e). We'll briefly explain how these expected frequencies are obtained. Notice from Table 2.3 that 51.1 percent of the sample were willing to vote for a woman candidate, while 48.9 percent were not. If sex and voting preference are independent (i.e., not related), we should find the same percentages for males and females. In other words, 48.9 percent (or 176) of the males and 48.9 percent (or 264) of the females would be unwilling to vote for a woman candidate. (This explanation is adapted from Norusis, 2011.) Now, we want to compare these two sets of frequencies to see if the observed frequencies are really like the expected frequencies. All we do is to subtract the expected from the observed frequencies (column three). We are interested in the sum of these differences for all cells in the table. Since they always sum to zero, we square the differences (column four) to get positive numbers. Finally, we divide this squared difference by the expected frequency (column five). (Don't worry about why we do this. The reasons are technical and don't add to your understanding.) The sum of column five (12.52) is called the **chi square statistic**. If the observed and the expected frequencies are identical (no difference), chi square will be zero. The greater the difference between the observed and expected frequencies, the larger the chi square.

If we get a large chi square, we are willing to reject the null hypothesis. How large does the chi square have to be? We reject the null hypothesis of no relationship between the two variables when the probability of getting a chi square this large or larger by chance is so small that the null hypothesis is very unlikely to be true. That is, if a chi square this large would rarely occur by chance (usually less than once in a hundred or less than five times in a hundred). In this example, the probability of getting a chi square as large as 12.52 or larger by chance is less than one in a thousand. This is so unlikely that we reject the null hypothesis, and we conclude that the alternative hypothesis (i.e., there is a relationship between sex and voting preference) is credible (not that it is necessarily true, but that it is credible). There is always a small chance that the null hypothesis is true even when we decide to reject it. In other words, we can never be sure that it is false. We can only conclude that there is little chance that it is true.

Just because we have concluded that there is a relationship between sex and voting preference does not mean that it is a strong relationship. It might be a moderate or even a weak relationship. There are many statistics that measure the strength of the relationship between two variables. Chi square is not a measure of the strength of the relationship. It just helps us decide if there is a basis for saying a relationship exists regardless of its strength. **Measures of association** estimate the strength of the relationship and are often used with chi square. (See Appendix D for a discussion of how to compute the two measures of association discussed below.)

Cramer's V is a measure of association appropriate when one or both of the variables consists of unordered categories. For example, race (white, African American, other) or religion (Protestant, Catholic, Jewish, other, none) are variables with unordered categories. Cramer's V is a measure based on chi square. It ranges from zero to one. The closer to zero, the weaker the relationship; the closer to one, the stronger the relationship.

Gamma (sometimes referred to as Goodman and Kruskal's Gamma) is a measure of association appropriate when both of the variables consist of ordered categories. For example, if respondents answer that they strongly agree, agree, disagree, or strongly disagree with a statement, their responses are ordered. Similarly, if we group age into categories such as under 30, 30 to 49, and 50 and over, these categories would be ordered. Ordered categories can logically be arranged in only two ways-low to high or high to low. Gamma ranges from zero to one, but can be positive or negative. The sign of Gamma is arbitrary since it depends on the way the categories are arranged. So it's best to ignore the sign and focus on the numerical value. You can use the percentages to decide on the direction (i.e., positive or negative) of the relationship. Like V, the closer to zero, the weaker the relationship and the closer to one, the stronger the relationship.

Choosing whether to use Cramer's V or Gamma depends on whether the categories of the variable are ordered or unordered. However, dichotomies (variables consisting of only two categories) may be treated as if they are ordered. For example, sex is a **dichotomy** consisting of the categories male and female. There are only two possible ways to order sex-male, female and female, male. Or, race may be classified into two categories-white and nonwhite. We can treat dichotomies as if they consisted of ordered categories because they can be ordered in only two ways. In other words, when one of the variables is a dichotomy, treat this variable as if it were ordered and use gamma. This is important when choosing an appropriate measure of association.

In this chapter we have described how surveys are done and how we analyze the relationship between two variables. In the next chapter we will explore how to introduce additional variables into the analysis.

REFERENCES AND SUGGESTED READING

Methods of Social Research

- Earl Babbie, 2009, *The Practice of Social Research* (12th edition), Wadsworth.

- Kenneth R. Hoover and Todd Donovan, 2011, *The Elements of Social Scientific Thinking* (10th edition), Wadsworth.

Interviewing

- Gorden, Raymond L., 1987, *Interviewing: Strategy, Techniques and Tactics*, Dorsey.
- Gorden, Raymond L., 1992, *Basic Interviewing Skills*, Waveland Press.

Survey Research and Sampling

- Earl Babbie, 2009, *The Practice of Social Research* (12th edition), Wadsworth.
- Arlene Fink, Linda Bourque, Eve P. Fieldler, Sabrina Mertens Oishi, Mark S. Litwin, 2003, *The Survey Kit* (2nd edition), Sage.
- Don A. Dillman, Jolene D. Smyth and Leah Melani Christian, 2009, *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method* (3rd edition), Prentice-Hall.

Statistical Analysis

- Herman J. Loether and Donald G. McTavish, 1993, *Descriptive and Inferential Statistics An Introduction* (4th edition), Allyn and Bacon.
- J. Richard Kendrick, Jr., 2004, *Social Statistics: An Introduction using SPSS* (2nd edition), Mayfield.
- Marija J. Norusis, 2011, *IBM SPSS Statistics 19 Guide to Data Analysis*, Prentice Hall.