DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Research Methods Posc 302

RESEARCH DESIGN
ANALYSIS OF SURVEY DATA

I. TODAY'S SESSION:
   A. First steps in data analysis and interpretation

II. PROBLEM:
   A. As always we should begin with a substantive problem.
   B. Suppose, to return to an earlier example, you have been hired by Hillary Clinton's campaign manager to provide insights into public opinion. In particular, the campaign's strategists want to know
      1. Is there any mileage to be gain in stressing women's issues?
         i. More specifically, is there a "gender gap" not just in political party affiliation but also in attitudes?
      2. Are voters, men and women, rational?
         i. Do people vote consistently with their issue positions or do they vote for candidates **in spite** of their stands on issues.
      3. Your "mission should you care to accept it" is to conduct some background research on voting behavior in the 1996 election and offer suggestions.
         i. Other than this broad guideline you have to work on your own.
         ii. But the report is due in 24 hours.
         iii. The First Lady will want a briefing that reassures her that her campaign ads can safely "reason with voters" and that she might stress issues popular among women.

III. HYPOTHESES:
   A. Right off the bat you have to ask some tough questions:
      1. What does the term "gender gap" mean, especially in the context of political attitudes.
         i. As a rough approximation, you might decide that if men and women differ on average by more than 15 or 20 percent on two or more important issues, there is reason to believe an important difference between men and women exists.
         ii. It's likely, moreover, given all the press coverage, that women are more "liberal" than men.
            1) But you keep open the possibility that any observed differences may be due to some "third" factor.
            2) That is, women may seem different from men only because

a disproportionate number of them live at or below the poverty line.

a) The "feminization of poverty" is a term you recall from a women's study course.

2. Next you need a definition of "rational voting."
   i. What the campaign manager seems to mean is the "logical" relationship between attitudes and candidate preference:
      1) people holding liberal opinions should vote for liberal candidates.
      2) Or, those having liberal opinions should vote for the candidates **closest** to them on those issues.
         a) Needless to say, conservatives should display similar consistency.
      3) People without opinions–the "apolitical or not-interested" citizens–will vote on other grounds, which you can't predict in this study.
   ii. To limit the study's scope you decide to use **presidential** voting in 1996 as a dependent variable.
3. Since the terms "liberal" and "conservative" have come up, you need to have a clear idea about their meaning.
   i. For now why not work with the conventional or common view: liberals favor an active government and public welfare programs while conservatives claim to want less government activity and more private initiative. In addition, liberals more than conservatives would favor less defense spending.

B. Summary: the main research hypotheses are thus:
1. There is a gender gap.
   i. There will be a substantial relationship between gender (X) and attitudes on major public policy controversies (Y).
   ii. There will be a relationship between gender (X) and liberalism-conservatism (Y): women will be more liberal.
2. Voters will vote rationally:
   i. In the 1996 presidential election conservatives will support conservative candidates; liberals will support liberals; and independents will vote (if they vote) on other grounds.

C. These hypotheses suggest the kinds of data we need to look for.

IV. DESIGN:

A. Given your time and resources limits you decide to conduct a **secondary analysis** of a major election study, the 1996 American National Election Study carried out by the Center for Political Studies at the University of Michigan.
1. You will adapt your hypotheses to the data collected in that study.

2. That is, you may have to modify definitions of say liberal and conservative in light of the questions asked in the 1996 survey.
3. Moreover, your choice of issues will be limited to the subjects covered in that study.
4. Still, experience tells you that there is plenty of information to help you answer the Clinton campaign's questions.

B. You can most easily test the hypotheses with cross-tabulations.
1. Before starting draw or sketch what you are looking for.
2. Example: you'll want tables that look something like this:

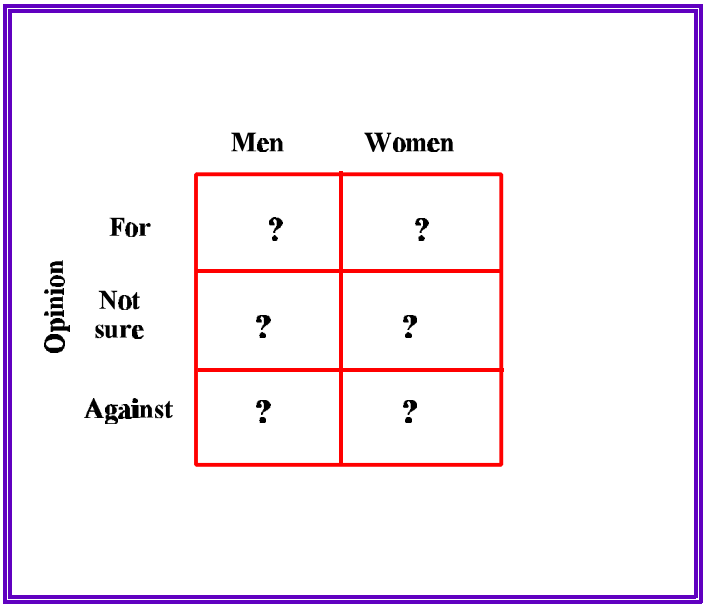


Figure 1: Plan Ahead - What Will the Analysis Look Like?

V. MEASUREMENT:

A. In a "normal" research project you would devise measures of key concepts such as liberal and conservative yourself.
1. But since you don't have the resources to conduct your own poll, you need to find out what measurements the 1996 Election Study has made.
2. To do that proceed as follows.

B. Use a desk top Windows PC that has access to the Internet.
1. Open "Notepad," the small word processor that comes with Windows, so you can take notes.
2. Open the browser, Internet Explorer or Netscape or another one and go to SDA: Survey Documentation and Analysis page at the University of California:

http://csa.berkeley.edu:7502/

i. It may be easiest to go to the course web site (www.udel.edu/htr/Researrchmethods) and click on "Sources of Information," "Raw Data," and finally "SDA."

3. Click on "SDA Archive."
4. Click on "(NES for 1996 (version with new weights) "
   i. You don't want the abstract; you wan the full study.
   ii. The address is: http://csa.berkeley.edu:7502/cgi-bin12/hsda?harcsda+nes96

C. The questions are describe in documentation called a codebook.
   1. Make sure that the circle next to "Browse codebook in this window" is checked and then click "Start."

D. You can then look for questions, which here are called **variables** in several ways.
   1. The easiest is to use "Headings for Groups of Variables"
   2. The figure on the next page shows part of the list of headings.
      i. The "Headings" list names groups of variables. To find out what's in a group just click on it.

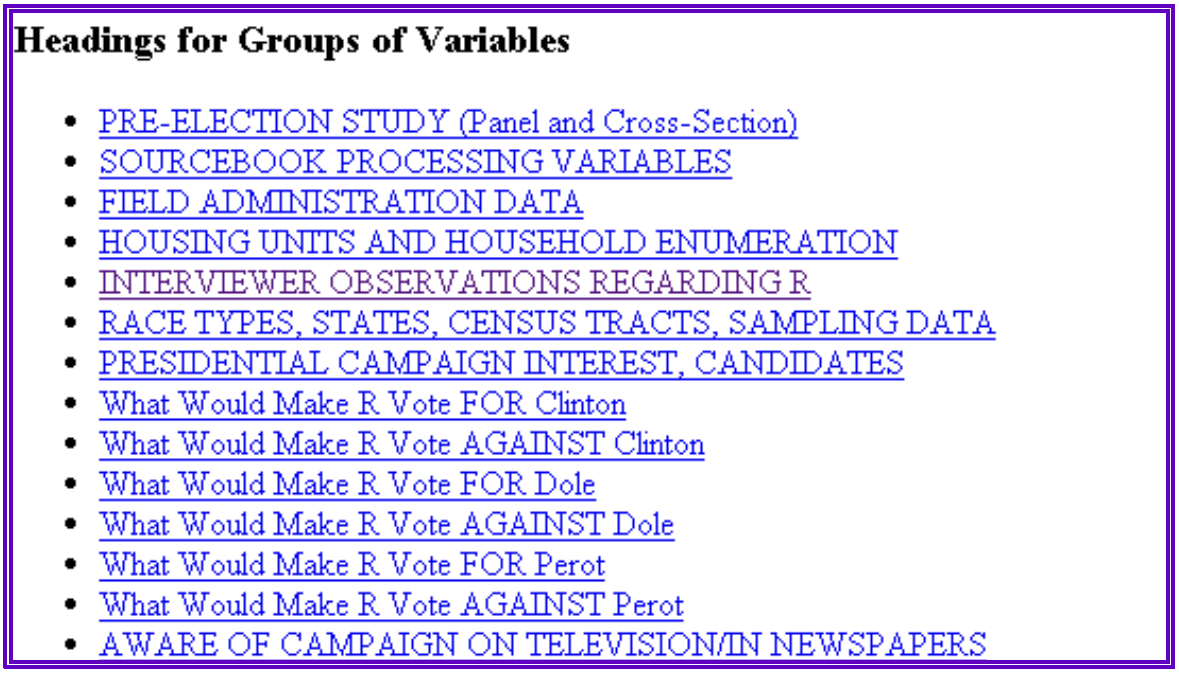


Figure 2: Groups of Variables in the 1996 American National Election Study

3. You need to know the respondents' sex so click on "INTERVIEWER

OBSERVATIONS REGARDING R ," which contains variables such as gender and race.

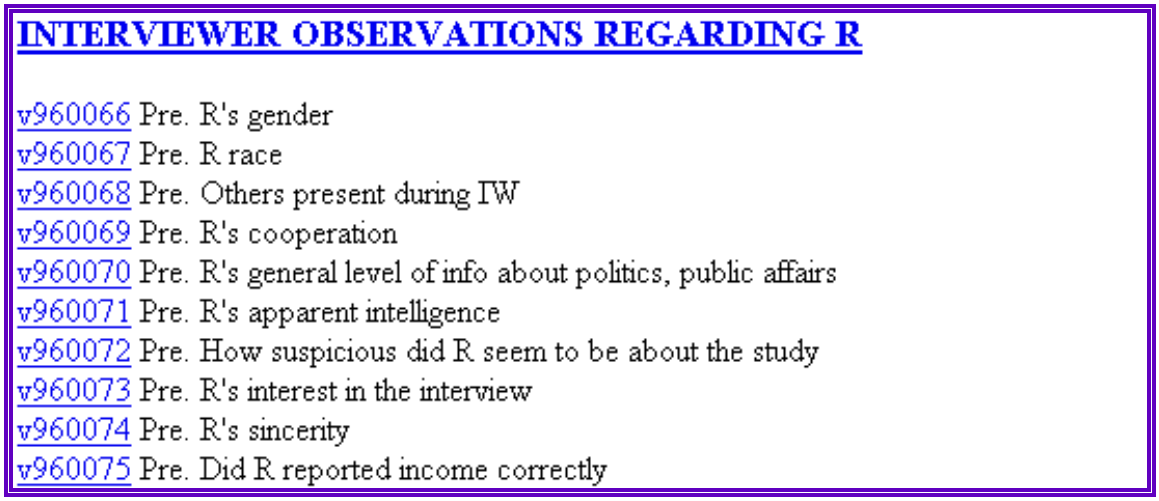


Figure 3: Variable List

4. Note the first line in the previous figure: v960066 Pre. R's gender
   i. It consists of two parts, a **variable** number and a brief variable **label**.
      1) The label of course tells you roughly the contents of the question or variable but you need to see the whole thing to make sure its what's needed for the analysis.
      2) Thus click on v960066 to see the text of the item.
      3) Copy and paste it to Notepad.

```
 v960066            Pre. R's gender


CSheet.31

PRE-ELECTION COVERSHEET
INTERVIEWER'S SUPPLEMENT, Item Z1.

R's sex is:

-----------------------------------------------------
-----------

    PCT     PCT      N  VALUE  LABEL
  VALID     ALL
   44.8    44.8    768      1  Male
   55.2    55.2    946      2  Female
                  ----
                 1714 cases
```

Text Box 1: Codebook Entry

E. We've seen this kind of information before, but this time we need to pay attention.
   1. The letter **R** means "respondent"
   2. All of the variables (questions) in the study have a number.
      i. The number is used in the analysis procedures so you need to pay careful attention to it.
         1) The variable number for gender is v960066.
         2) Here's what it means (in this study) (See the next figure.)

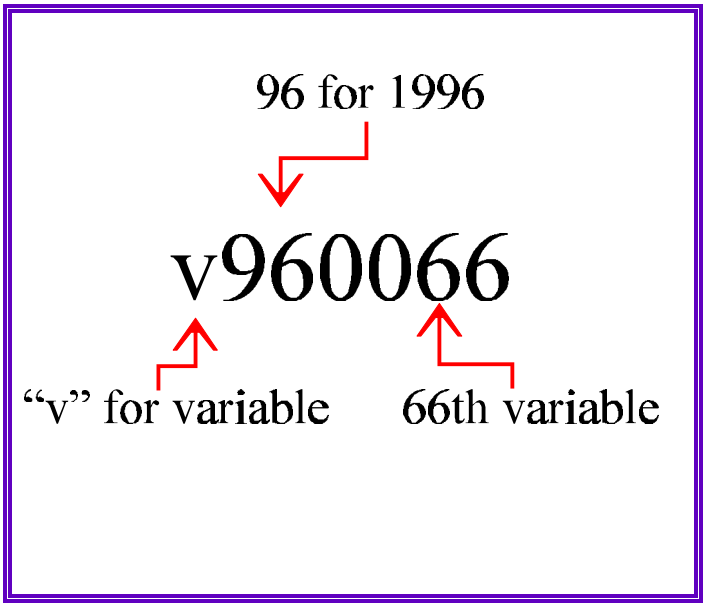


Figure 4: Variable Numbers - Pay Attention!

ii. Knowing that there is a method in the numbering madness may help you keep track of the numbers.

iii. **In any event these number must be entered exactly as they appear in the codebook when using the analysis procedures in the SDA system.**

3. Value **labels**
   i. The variable gender has two **values** labeled "Male" and "Female"
   ii. These names or labels have been assigned codes so that a computer can easily process the information.
   iii. The code or value for males is "1." If you see a one on an analysis table using this variable, it means "males."
      1) Similarly, females are coded "2.."
   iv. The values or codes are arbitrary. In some studies men are coded "5" and women "1." It doesn't matter just as long as the codes or labels are fully described by the codebook.
   v. Later we will change these values or codes.
4. The total number of people interviewed in the poll is indicated in the "cases" line.
   i. The 1996 National Election Study contains 1714 respondents..
5. Frequencies:
   i. The codebook also tells you how many people fall into each category or code. (Look in the column labeled "N.")
      1) There are 768 males and 946 females.
      2) Everyone in this study fell into one of the two categories of

the variable.

6. Percentages
   i. Under the "all" column one finds the percentages of the total 1714 respondent who fell into each category.
      1) We see that 44.8 percent of the 1714 respondents were male. That's (768/1714)100 = 44.8.
      2) Similarly 55.2 percent of the respondents are female.
      3) The percentages in the "All" category add to 100.
   ii. The percent "valid" raises an important point but let's look at another example to see it.

F. Besides gender we need to know how these respondents felt about issues. So go back to the headings list and scroll down.
1. There are lots and lots of attitudinal questions so to limit our study let's choose health care and spending on defense, two issues that are likely to be important in the 2000 campaign.
2. About a quarter of the way down find a topic labeled "GOVERNMENT HEALTH INSURANCE SCALE."
   i. Click on it and go to variable "v960479 Pre. R's self-placement on govt health insurance scale."
      1) Note: we want to know what R felt about an issue, not where R would place a candidate.
      2) Moral: read the item labels carefully and if in doubt read the entire question.
   ii. Figure 5 below shows a part of the codebook entry.
3. In particular it shows the "valid" responses and the codes or values assigned to them.
   i. This is a seven-point scale on which "1" means fully supports a government health insurance program and 7 indicates a preference for a private system.
   ii. Scores of 3, 4, and 5 represent the opinions of respondents who are some where in the middle on this issue.
4. Note that on this issue like so many others some respondents don't have a preference.
   i. They are put in the "Haven't though much about it" category and scored 0.
   ii. For many purposes we can treat these individuals as if they didn't participate in the research and thus consider their answers "missing" or "invalid."
5. By the same token some people did not respond at all or their answers were not recorded. They too are consider "missing."
6. The codebook distinguishes between substantive and missing or invalid responses.

i. Thus it calculates two percentages:
   1) Percent of all people in the study (1714 in all in this case) and
   2) Percent of those who gave "'valid" or substantive responses.
   3) See Figure 5.

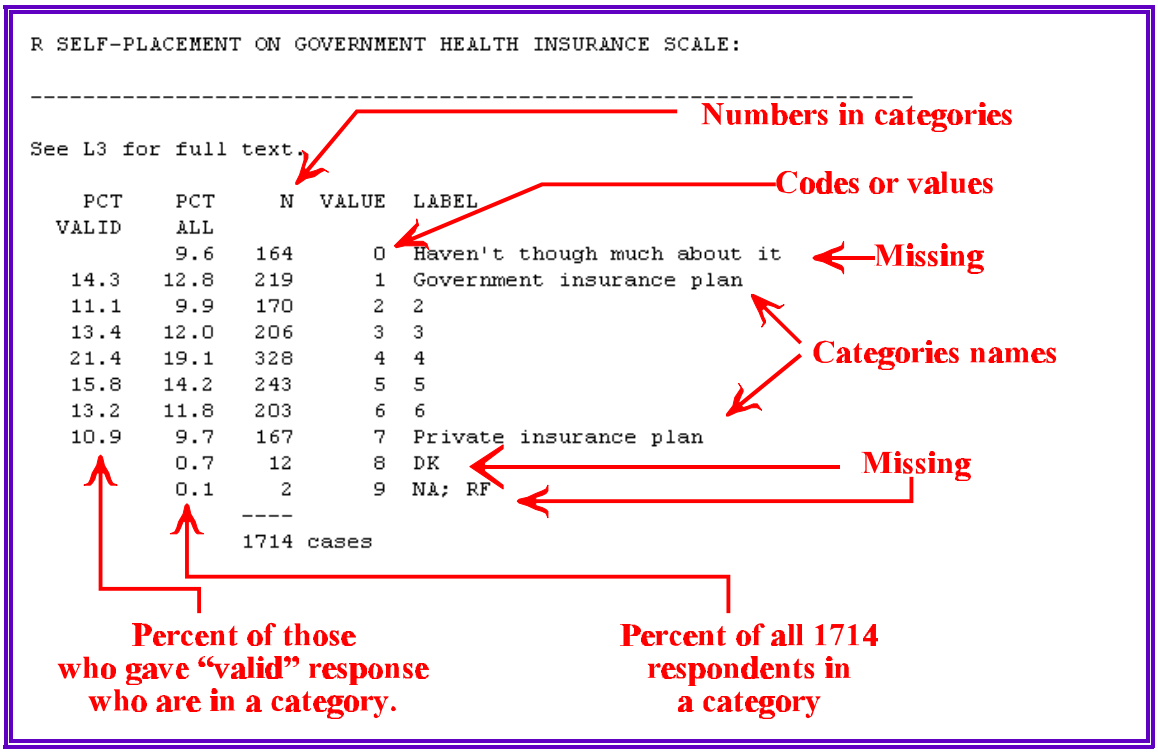


Figure 5: Codebook Entry with Missing Data Indicated

   4) Example: 219 out of 1714 total cases or 12.8 percent favor government health insurance.
   5) But these 219 constitute 14.3 percent of those who gave valid or non-missing responses.

G. **Note: as a general rule we will eliminate missing data from our analyses.**
   1. This is a tricky proposition and will return to it time and again. But it is standard in academic research.

H. As noted above you will include attitudes toward defense spending as well.
   1. Found under the "DEFENSE SPENDING SCALE" heading it has the same form as government health insurance and we'll analyze in a similar fashion.

I. Finally, we need a measure of how people actually voted in 1996.
   1. Scroll about two thirds of the way down the Headings page to find" VOTE

FOR PRESIDENT/HOUSE/SENATE CANDIDATE."

i. Click on it and then "v961082 Post. Which presidential candidate did R vote for."

ii. Copy and paste the entry to Notepad.

```
v961082            Post. Which presidential candidate did R vote for


D8

IF R VOTED:
IF R VOTED FOR CANDIDATE FOR PRESIDENT:
Who did you vote for?

-------------------------------------------------------------------------
-

    PCT     PCT      N  VALUE  LABEL
  VALID     ALL
           32.3   554      0  Inap, R did not vote (5,6,8,9 in D1); R
                              voted but not for  President
   52.9    35.0   600      1  Bill Clinton
   38.3    25.3   434      2  Bob Dole
    7.2     4.8    82      3  Ross Perot
    1.6     1.1    18      7  Other
            1.5    25      8  Refused
            0.1     1      9  NA
                 ----
                1714 cases
```

Text Box 2: Vote For President Codebook Entry

iii. 554 (about a third of all the respondents) did not vote in the 1996 presidential election.

1) They are counted as missing or invalid response..

2) So too are the 25 people who refused to answer and the 1 person for who information was not ascertained.

iv. Note that 600 people claimed to have voted for Clinton.

1) This number represents 35.0 percent of all the respondents but 52.9 percent of those who cast a ballot of one kind or another.

2. You should restrict your analysis to valid responses.

J. Of course you could choose to examine other variables.

1. But look at the list. Unless you start with a plan, which is embodied in a set

of hypotheses, you'll soon get lost.

VI. ANALYSIS:

A. Now that you have established hypotheses and measurements you can "test" your hypotheses.

B. Method:

1. A simple first step is to note marginal totals: these are simply the frequencies and percentages listed in the codebook.
   i. Keep in mind both the total and "valid" percentages.
   ii. We saw, for example, that about 30 percent of the respondents in the 1996 National Election Study did not vote. So any conclusions about "rational voting" and using issue-oriented campaigns will have to keep that point in mind.
   iii. It's important to study the other marginal totals before proceeding.
2. Next think about relationships.
   i. The initial plan was to create cross-tabulations in order to investigate the hypotheses.

C. Creating cross-tabs:

1. I suggest that you open a second browser window and go to the SDA page, then to the 1996 American National Election study as before.
2. This time check "Frequencies or cross tabulation" and click start.
3. You'll see a dialog box like the one shown in the next figure.

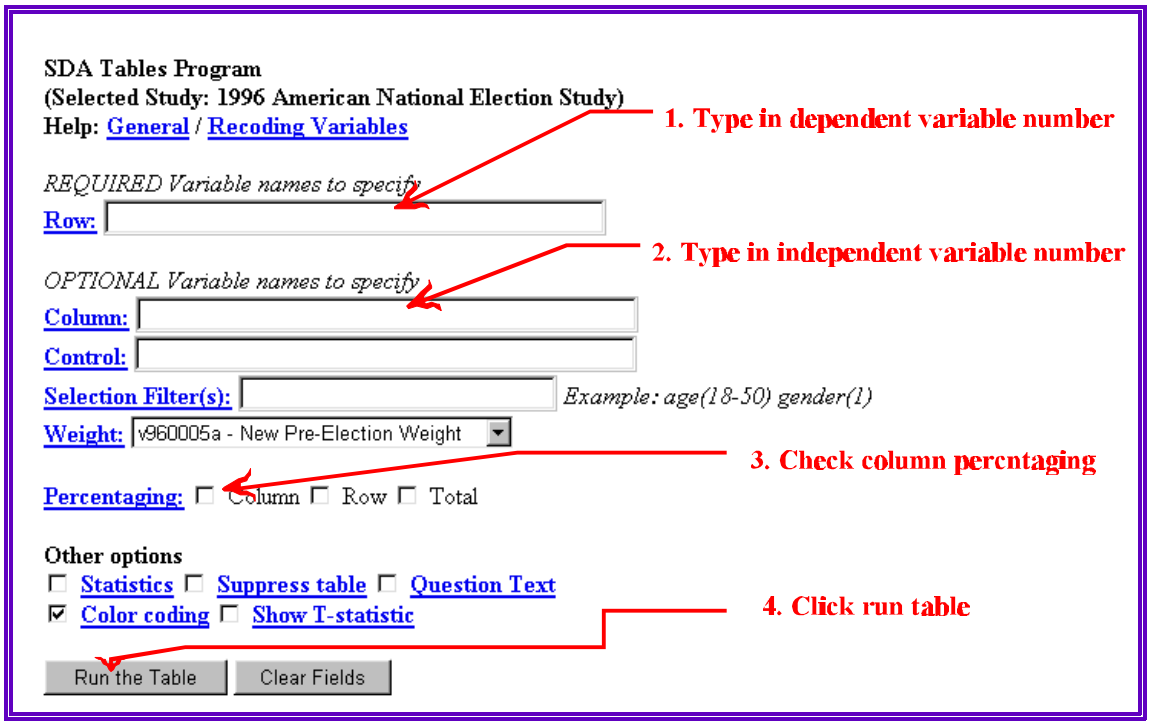


Figure 6: Cross-tabulation Dialog Box

4. It's simple to make a cross-tabulation or contingency table if you take your time and make sure variable names are correctly typed and enter in the proper place.
   i. I prefer that the dependent variable be the row variable and the independent variable is the column.
   ii. It's also going to be important to obtain column percentages in order to simplify the interpretation of the data.
   iii. To avoid mistakes I sometimes use cut and past **variable numbers only** from Notepad to the appropriate dialog box entry.
5. Example.

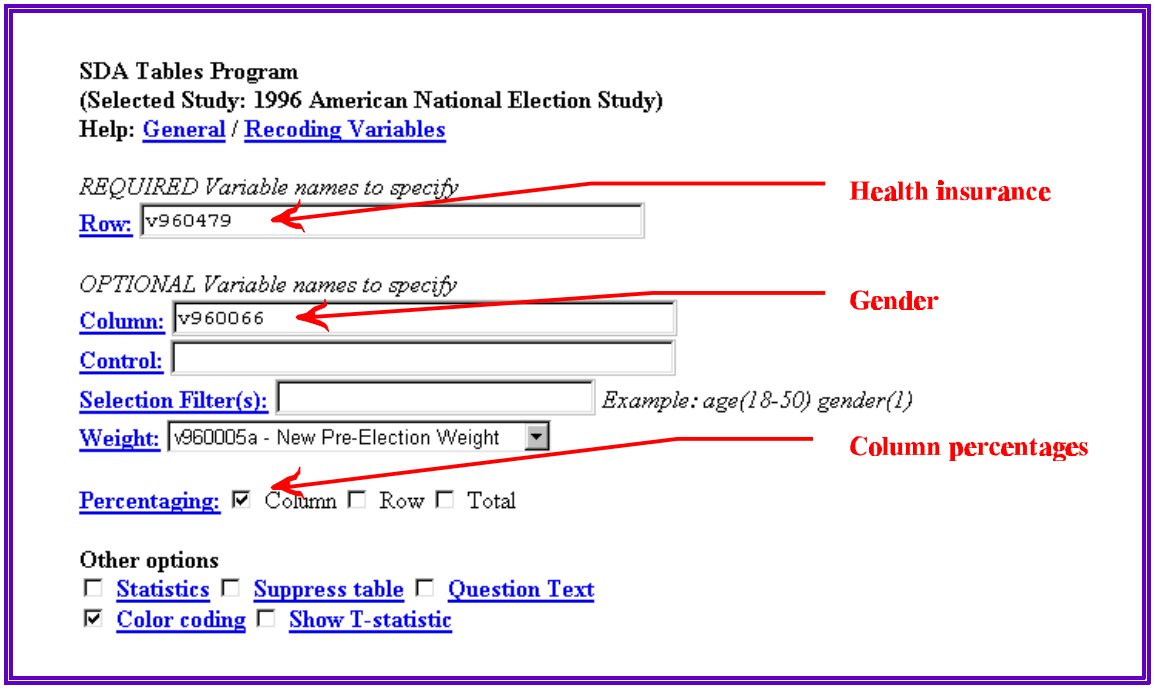
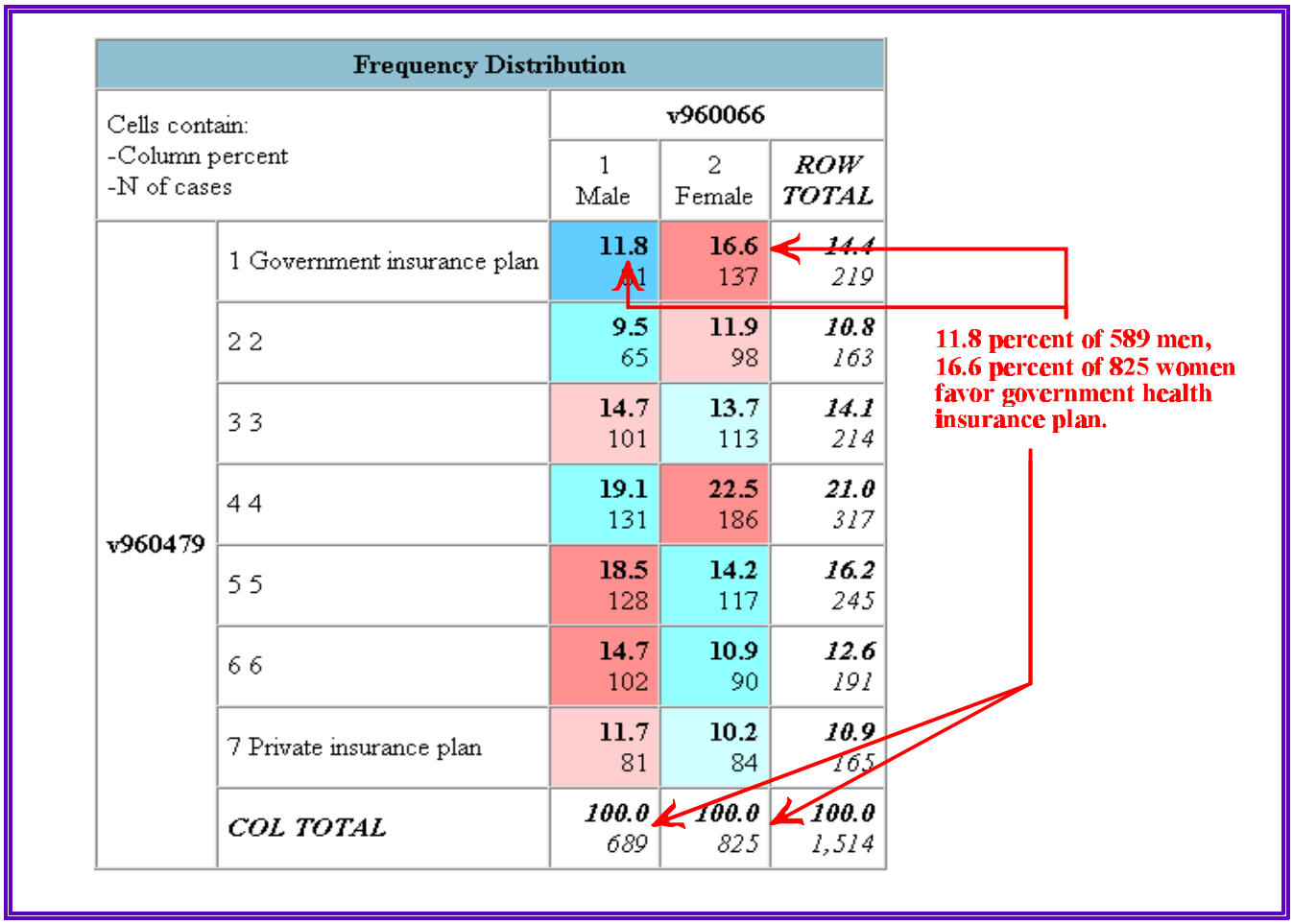


Figure 7: Cross-tabulation of Attitude on Health Insurance by Gender

6. This particular specification produces the following table.
   i. See Figure 8 on the next page.
   ii. It suggests that there are differences between men and women on this issue.



Frequency Distribution				
Cells contain: -Column percent -N of cases		v960066		
		1 Male	2 Female	*ROW TOTAL*
v960479	1 Government insurance plan	**11.8** 81	**16.6** 137	***14.4*** *219*
	2 2	**9.5** 65	**11.9** 98	***10.8*** *163*
	3 3	**14.7** 101	**13.7** 113	***14.1*** *214*
	4 4	**19.1** 131	**22.5** 186	***21.0*** *317*
	5 5	**18.5** 128	**14.2** 117	***16.2*** *245*
	6 6	**14.7** 102	**10.9** 90	***12.6*** *191*
	7 Private insurance plan	**11.7** 81	**10.2** 84	***10.9*** *165*
	COL TOTAL	***100.0*** *689*	***100.0*** *825*	***100.0*** *1,514*

11.8 percent of 589 men, 16.6 percent of 825 women favor government health insurance plan.

Figure 8: Attitudes Toward Health Insurance By Gender

iii. We'll discuss the interpretation of the table next time.

VII. TEST YOURSELF:

A. Suppose you are asked to work for the George Bush campaign, which wants to know if a gender gap exists on foreign policy issues.

1. Go to the variable headings list of the 1996 American National Election study.
2. Find as many foreign policy issues as you can.
3. Make sure that you understand the code book entries.
   i. How many and what percent of **all** the people favor various alternatives?
   ii. How of those giving **valid** or **non-missing responses** favor the alternatives?
   iii. Is there a difference between i. and ii?
4. Try creating a cross tabulation table between gender and one of the foreign policy issue items you selected.
   i. The numbers in the table are frequencies. You should be able to calculate the percentages **and** row and column marginal totals.
   ii. How many total cases are in this table?
   iii. Would you say that there is a relationship between region (X) and attitudes toward gun control? Explain.

VIII. NEXT TIME:

A. Survey research

1. We start analyzing survey data by creating simple cross-tabulations.
2. Multi-way tables

B. Reading:

1. Johnson and Joslyn, ***Research Methods***, Chapter 12, pages 327 to 336 describes cross-tabulations.
   i. Table 12.1 shows how these tables are constructed from codes.
   ii. Generally the authors follow the format I prefer by making the independent variable the column variable. Note, however, that not everyone follows this convention.
   iii. Tables 12.7 and 12.8 illustrate the notion of "strength of relationship" and may help you interpret your results.