PSCI 241: American Public Opinion and Voting Behavior
Statistical Analysis of the 2000 National Election Study in STATA

Introduction

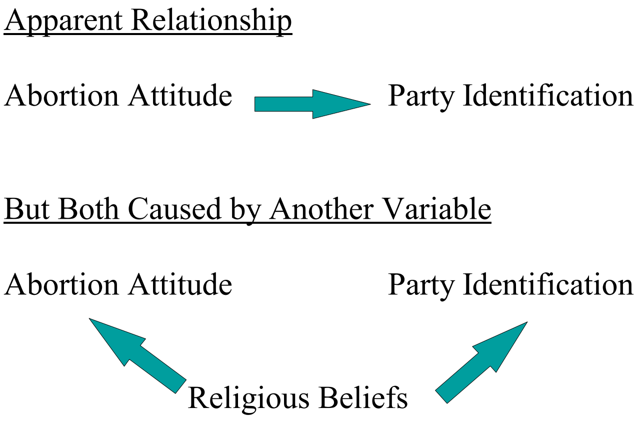
This document explains how to work with data from the 2000 National Election Study (NES) and perform statistical analysis on that data in the statistical software program STATA. All of the examples are based on the following research questions: what is the relationship between attitudes on the issue of abortion and political behavior? Some hypotheses would be: (1) Individuals with pro-life attitudes on abortion are more likely than individuals with pro-choice attitudes to identify with the Republican party. (2) Individuals with pro-life attitudes on abortion are more likely than individuals with pro-choice attitudes to feel positively toward Republican presidential candidates and negatively toward Democratic presidential candidates. (3) Individuals with pro-life attitudes on abortion are more likely than individuals with pro-choice attitudes to vote for Republican presidential candidates. So, the primary independent variable in this analysis will be attitude on the abortion issue. The primary dependent variables will be party identification, comparative candidate evaluations (measured as the feeling thermometer rating of George Bush minus the feeling thermometer rating of Al Gore) and the 2000 presidential vote.

Choosing Variables from the 2000 NES Codebook

The first step in testing these hypotheses is to select the variables from the 2000 NES that will be necessary to adequately test them. The three types of variables we will need to conduct the appropriate tests of our hypotheses are the **independent variables**, the **dependent variables**, and **the control variables**. We already have identified the main independent variable (abortion attitudes) and dependent variables (party identification, comparative candidate evaluations, and the presidential vote) in our analyses. All we have to do now is to figure out how to **operationalize** those variables, i.e. figure out how to measure them using the 2000 NES data. So, the main thing to do at this point is to figure out which variables we should use as control variables and decide on how to operationalize those variables.

Control variables are variables that may affect or explain the relationship–the way in which change in one variable is associated with change in another variable–between the independent and dependent variables. There are two ways in which other variables may affect that relationship. One way is the case of a **spurious** relationship between the independent and dependent variables. The relationship between two variables is spurious if what appears to be a relationship between the two is actually due to the fact that both variables are caused by some other variable. In other words, the reason that changes in an independent variable are associated with changes in a dependent variable is because both the changes in both variables result from changes in some other variable. Suppose, for example, that we observe a relationship between abortion attitudes and party identification: as individuals grow more pro-life on abortion, they become more likely to identify themselves as Republicans. Perhaps that relationship is spurious because changes in abortion attitudes and in party identification may result from changes in religious beliefs. Individuals with orthodox religious beliefs are more likely than individuals with progressive religious beliefs to have pro-life attitudes on abortion, and individuals with orthodox religious beliefs are more likely than individuals with progressive religious beliefs to identify with the Republican party.

Figure 1: A Potentially Spurious Relationship Between Abortion Attitude and Party ID



Apparent Relationship

But Both Caused by Another Variable

To see if abortion attitudes really are related to party identification, or if that relationship is spurious due to the two variables' mutual relationship with religious beliefs, we need to **control** for religious beliefs. In other words, we need to examine the relationship between abortion attitudes and party identification, while **holding** religious beliefs **constant**: holding them at the same value so that any observed relationship between changes in abortion attitudes and changes in party identification cannot be due to changes in religious beliefs. If we still observe a relationship between abortion attitudes and party identification while controlling for (or holding constant) religious beliefs, then we may conclude that their relationship is not spurious. If we no longer observe a relationship between abortion attitudes and party identification while controlling for religious beliefs, then we must conclude that their relationship is spurious.

Another way in which another variable can affect or explain the relationship between an independent variable and a dependent variable is in the case of an **intervening** relationship: when some other variable intervenes between the independent and dependent variables, explaining why they are related. For example, perhaps the reason that abortion attitudes are related to party identification is that attitude on abortion affects more general ideological orientations, or the extent to which one considers oneself a liberal or conservative, and those ideological orientations in turn affect party identification. In other words, abortion attitudes do affect party identification, but rather than a **direct** effect, the effect is **indirect**.

Figure 2: An Indirect Relationship Between Abortion Attitude and Party ID

Abortion Attitude → Liberal-Conservative Identification → Party ID

To determine whether abortion attitude has a direct or an indirect effect on party identification, we need to examine the relationship between those two variables while controlling for liberal-conservative identification (i.e. hold it constant so that an observed relationship between changes in abortion attitude and changes in party identification cannot be due to changes in liberal-conservative identification). If we still observe a relationship between abortion attitude and party identification while controlling for liberal-conservative identification (and the other variables that may intervene between abortion attitude and party identification), then we can conclude that abortion attitude has a direct effect on party identification. If we no longer observe a relationship between abortion attitude and party identification while controlling for liberal-conservative identification, we must conclude that abortion attitude has an indirect effect on party identification, that the relationship between abortion attitude and party identification is explained by liberal-conservative identification.

So, what we need to do is to try to identify the variables for which we need to control in order to assess the nature of the relationship between abortion attitude and party identification (or comparative candidate evaluations or the presidential vote). That includes the variables that may cause both abortion attitude and party identification (producing a spurious relationship between the two) and the variables that may intervene between abortion attitude and party identification. In short, we should control for the variables that we think will be related to both abortion attitude and party identification. This list of variables should be based on our own common sense knowledge of politics and our reading of the scholarly literature on abortion attitudes and political behavior. Such a list would include demographic and religious factors that may shape both abortion attitude and party identification, attitudes toward other political issues that may be related to both, and more general political orientations (such as liberal-conservative identification) that may be related to both. Suppose we come up with the following list.

Figure 3: Control Variables for Examining the Relationship Between Abortion Attitude and Party Identification

Demographic and Religious Variables	Attitudes on Other Political Issues	General Political Orientations
Education Income Gender Region of Residence (South or Non-South) Religious Beliefs (View of the Bible) Worship Attendance Race Age	Attitudes on Other Cultural Issues (Homosexual Discrimination Laws, Women's Equal Rights) Attitudes on Other Types of Issues (Defense Spending, Government Guarantee of Jobs, Government Help for African-Americans)	Ideology (Liberal-Conservative Identification)

The next step is to go to the codebook for the 2000 NES and find these control variables, the independent and dependent variables, and the respondent ID number (necessary to merge new data into your existing data set) and the relevant information for them. We first need to look at the "variable description list" in the codebook and find the variable numbers for these variables. That yields the following list in Figure 4.

Figure 4: Variable Numbers for Relevant Variables from the 2000 NES

Variable	Number
Case ID	v000001
Education	v000913 (summary measure)
Income	v000977 (household income)
Gender	v001029 (interviewer's observation)
Region	v00092 (census region)
View of Bible	v000876
Worship Attendance	v000877, v000879, v000880 (need to combine into one variable in STATA)
Race	v001030 (interviewer's observation)
Age	v000908
Party Identification	v000523 (summary measure)
Abortion Attitude	v000694

2000 Presidential Vote	v001249 (post-election report of vote)
Homosexual Discrimination Laws	v001481 (summary measure)
Women’s Rights	v000760 (combined 7-point and branching measures)
Defense Spending	v000587 (combined 7-point and branching measures)
Government Guarantee Jobs	v000620 (combined 7-point and branching measures)
Government Help for Blacks	v000645 (combined 7-point and branching measures)
Liberal-Conservative Identification	v000440 (just 7-point scale respondents)
Gore Feeling Thermometer	v000360
Bush Feeling Thermometer	v000361

Once we find the numbers of our variables, we then go to the “variable documentation” section of the codebook and find out the relevant information about our variables: the wording of the questions and the values corresponding to various responses. For example, when we go to the documentation on worship attendance, we find that there are three questions that are relevant: v000877, v000879, v000880. The documentation for these questions is as follows:

```
==============================
VAR 000877 X1. Attend religious services
MD1: EQ 0, MD2: GE 8
COLUMNS: 1772 - 1772
Numeric
X1.
Lots of things come up that keep people from attending
religious services even if they want to. Thinking about
your life these days, do you ever attend religious
services, apart from occasional weddings, baptisms or
funerals?
----------------------------------------------------------------------
1. YES --> SKIP TO X2
5. NO --> SKIP TO X1a
8. DK --> SKIP TO X1a
9. RF
0. NA
0 1 5 9
----- ----- ----- -----
```

```
==============================
VAR 000879 X2. Attend religious services how often
MD1: EQ 0, MD2: GE 8
COLUMNS: 1774 - 1774
Numeric
X2.
IF R ATTENDS RELIGIOUS SERVICES:
Do you go to religious services every week, almost every
week, once or twice a month, a few times a year, or
never?
--------------------------------------------------------------------
1. EVERY WEEK --> X2a
2. ALMOST EVERY WEEK --> X3
3. ONCE OR TWICE A MONTH --> X3
4. A FEW TIMES A YEAR --> X3
5. NEVER --> X3
8. DK --> X3
9. RF
0. NA; INAP, 0,5,8,9 in X1

==============================
VAR 000880 X2a. Attend relig serv > once/week
MD1: EQ 0, MD2: GE 8
COLUMNS: 1775 - 1775
Numeric
X2a.
IF R SAYS ATTENDS RELIGIOUS SERVICES 'EVERY WEEK':
Would you say you go to religious services once a week
or more often than once a week?
--------------------------------------------------------------------
1. ONCE A WEEK
2. MORE OFTEN THAN ONCE A WEEK
8. DK
9. RF
0. NA; INAP, 5,8,9, 0 in X1; 2-5,8,9 or NA in X2
```

So, if respondents answered “no” to the first question (v000877), they were not asked the second question (v000879). If they answered “yes” to the first question, they were asked the second question. Then, if respondents answered “every week” to the second question, they (and only they) are asked a third question (v000880). To form a measure of worship attendance ranging from “never attend” to “attend more often than once a week,” we will have to combine the responses to these three questions in STATA.

Once we download the relevant variables from the 2000 NES (I will do that for you), we are ready to begin working with the data in STATA.

Opening and Saving Data in STATA

I will email each of you a STATA data file named "nes2000_yourname.dta". When you receive the email from me, you should right-click on the attachment with your mouse and choose "save." Then save the data file to either your hard drive or a disk. Depending on how many variables you want in your data set, the file may be too large to fit on a normal floppy disk. If so, you can either save the file to your hard drive and zip it (using WinZip or some such program) so that it will fit on a floppy, or save it to a zip disk or your hard drive. For the purposes of this example, let's assume for now that each of you is named "ps241." I would then email you a STATA data file named "nes2000_ps241.dta" and you are ready to begin manipulating and analyzing your data in STATA.

To open your data file in STATA, simply go to the file menu and click on **open**. You can then browse the hard drive or a disk for your data file. Click on your file and it will open in STATA. I doubt this will happen, but if you have a large data set, you may get an error message saying "no room to add more observations." If that happens, it means that there is not enough memory on the computer allocated to STATA for it to handle your data set. There is a simple solution: just increase the amount of memory allocated to STATA. The default in most labs on campus is 1 megabyte of memory allocated to STATA. If you increase it to 8 megabytes, you should be fine. You can do that simply by typing:

```
set mem 8m
```

Once the data is in Stata, we can save it using the **save as** command from the file menu. If you wish to save a file that you have saved before under the same name, just use the **save** command and indicate that you wish to overwrite the existing file, or simply type

```
save, replace
```

Stata automatically adds the suffix **.dta** to Stata-format data sets. Once you have saved the file and exited Stata, you can bring the file back into Stata with the **open** command from the file menu.

Before we get too far along, here are **three useful hints for using STATA**:

(1) Never use upper-case letters when typing Stata commands.
(2) If you want to rerun a previous command, you don't have to retype it. Just go back to it using the **page-up** key or scroll in the review window and click on the old command.
(3) If you make a mistake in a data set (e.g. delete a variable you wanted to keep, made a coding mistake in a variable, etc.), you should:

(a) Not save the data

(b) Reopen the data set. Since you made changes to the data and did not save it, Stata will ask you if you want to clear the current data from memory. Say yes.

Viewing Your Variables in STATA

Once you have opened your data file, the first thing you will probably want to do is see a list of the variables in your data set. You can do this by simply typing **d** (for describe). That shows us a list of the variables in our data set. The other thing that is relevant in this description of our variables is the “variable label.” The NES has been kind enough to provide us with labels for the variables we downloaded.

```
. d

Contains data from C:\PSCI 241\Fall 2002\nes2000_ps241.dta
  obs:         1,807
 vars:            22
 size:        52,403 (99.3% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
v000001         int    %8.0g                  process.4. case id
v000092         byte   %8.0g       v000092    pre.sample.15. census region
v000360         int    %8.0g       v000360    c1b/c1b.t. thermometer gore
v000361         int    %8.0g       v000361    c1c/c1c.t. thermometer george w
                                                bush
v000440         byte   %8.0g       v000440    g1ax. summary: combined ftf/ph
v000523         byte   %8.0g       v000523    k1x. party id summary
v000587         byte   %8.0g       v000587    l2ax2. comb.7pt/br summ defense
                                                spending
v000620         byte   %8.0g       v000620    l4x2. comb.7pt/br summ
                                                guaranteed jobs
v000645         byte   %8.0g       v000645    l5ax2. comb.7pt/br summ r aid
                                                to blacks
v000694         byte   %8.0g       v000694    m1/m1.t. abortion self-placement
v000760         byte   %8.0g       v000760    p1a1x2. comb.7pt/br summ r
                                                equal role
v000876         byte   %8.0g       v000876    s5/s5.t. bible is word of god
                                                or men
v000877         byte   %8.0g       v000877    x1. attend religious services
v000879         byte   %8.0g       v000879    x2. attend religious services
                                                how often
v000880         byte   %8.0g       v000880    x2a. attend relig serv >
                                                once/week
v000908         byte   %8.0g       v000908    y1x. respondent age
v000913         byte   %8.0g       v000913    y3x. r educ summary
v000994         byte   %8.0g       v000994    y27x. hh income -all hhs
v001029         byte   %8.0g       v001029    zz1. iwr obs: r gender
v001030         byte   %8.0g       v001030    zz2. ftf iwr obs: r race
v001249         byte   %8.0g       v001249    c6. r vote cast for president
v001481         byte   %8.0g       v001481    k11x. summary protctng homosxls
                                                against
-------------------------------------------------------------------------------
Sorted by:
```

Although we have these variable labels to tell us what each of our variables represent, our lives would be much easier if we had variable labels that were a bit more descriptive than v000001 and v001481. So, we might want to rename our variables using STATA’s **rename** command as follows:

rename v000001 caseid

rename v000876 bibview

rename v001249 presvote

If we renamed all of our variables (except the ones relating to worship attendance on which we still have some work to do), our data set would look like this:

```
. d

Contains data from C:\PSCI 241\Fall 2002\nes2000_ps241.dta
  obs:         1,807
 vars:            22
 size:        52,403 (99.3% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
caseid          int    %8.0g                  process.4. case id
region          byte   %8.0g       v000092    pre.sample.15. census region
goreft          int    %8.0g       v000360    c1b/c1b.t. thermometer gore
bushft          int    %8.0g       v000361    c1c/c1c.t. thermometer george w
                                                bush
ideology        byte   %8.0g       v000440    g1ax. summary: combined ftf/ph
partyid         byte   %8.0g       v000523    k1x. party id summary
defspend        byte   %8.0g       v000587    l2ax2. comb.7pt/br summ defense
                                                spending
govjobs         byte   %8.0g       v000620    l4x2. comb.7pt/br summ
                                                guaranteed jobs
helpblacks      byte   %8.0g       v000645    l5ax2. comb.7pt/br summ r aid
                                                to blacks
abortion        byte   %8.0g       v000694    m1/m1.t. abortion self-placement
womrights       byte   %8.0g       v000760    p1a1x2. comb.7pt/br summ r
                                                equal role
bibview         byte   %8.0g       v000876    s5/s5.t. bible is word of god
                                                or men
v000877         byte   %8.0g       v000877    x1. attend religious services
v000879         byte   %8.0g       v000879    x2. attend religious services
                                                how often
v000880         byte   %8.0g       v000880    x2a. attend relig serv >
                                                once/week
age             byte   %8.0g       v000908    y1x. respondent age
education       byte   %8.0g       v000913    y3x. r educ summary
income          byte   %8.0g       v000994    y27x. hh income -all hhs
sex             byte   %8.0g       v001029    zz1. iwr obs: r gender
race            byte   %8.0g       v001030    zz2. ftf iwr obs: r race
presvote        byte   %8.0g       v001249    c6. r vote cast for president
homdisc         byte   %8.0g       v001481    k11x. summary protctng homosxls
                                                against
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

We also might want to change some of the variable labels so that they are more descriptive. For example, the variable label for “ideology” does not tell us a whole lot. So, we might want to use STATA’s **label var** command to give it a new label:

```
label var ideology “7-point liberal-conservative identification”
```

If we then ask for a description of just that variable, we get the following:

```
. d ideology

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
ideology        byte   %8.0g       v000440    7-point liberal-conservative
                                                identification
```

Once we have seen the variables that are in our data set, the next thing we probably will want to do is take a look at the individual variables and see how the NES respondents are distributed across the various response options of those variables. In other words, we want to view a **frequency distribution** of the variable, which is a table of the outcomes, or response categories of the variable, and the number of times each outcome is observed. The **tabulate** or **tab** command in STATA produces a frequency distribution of a variable. Let’s take a look at the frequency distribution of abortion attitudes:

```
. tab abortion

       m1/m1.t. abortion self-placement |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
1. by law, abortion should never be per |        215       12.04       12.04
2. the law should permit abortion only  |        525       29.40       41.43
3. the law should permit abortion for r |        265       14.84       56.27
4. by law, a woman should always be abl |        753       42.16       98.43
               7. other (specify) [vol] |         28        1.57      100.00
----------------------------------------+-----------------------------------
                                  Total |       1786      100.00
```

The first column shows the various response options on the NES question about abortion: (1) by law, abortion should never be permitted, (2) the law should permit abortion only in the cases of rape, incest, or when the woman’s life is in danger, (3) the law should permit abortion for reasons other than rape, incest, or danger to the woman’s life but only when a clear need has been established, (4) by law, a woman should always be able to obtain an abortion as a matter of personal choice, and (7) a volunteered response that is something other than one of the NES response options. Unfortunately, the labels that the NES has provided for these response options do not do a great job of indicating what each one is. So, we might wish to come up with a new set of labels for these values that are more descriptive. We can do that with STATA’s **label define** and **label values** commands, as follows:

```
. label define abort 1 "never allow" 2 "rape/incest/life" 3 "other, clear need" 4 "always allow" 7 "other (vol.)"

. label values abortion abort
```

In the “label values” command, the variable for which we are labeling values (abortion) comes first, and the value label that you have defined using the “label define” command (abort) comes second.

If we then asked for a frequency distribution of the abortion variable, we get the following:

```
. tab abortion

m1/m1.t. abortion |
   self-placement |      Freq.     Percent        Cum.
------------------+-----------------------------------
      never allow |        215       12.04       12.04
 rape/incest/life |        525       29.40       41.43
other, clear need |        265       14.84       56.27
     always allow |        753       42.16       98.43
     other (vol.) |         28        1.57      100.00
------------------+-----------------------------------
            Total |       1786      100.00
```

The second column shows the frequency distribution for this variable – the number of respondents to the 2000 NES who chose the various response options to the abortion question. One thing to note is that there were 1,807 people who were surveyed for the 2000 NES, but only 1,786 total observations on this variable. That means that only 1,786 of the observations are **useable observations** – observations that are of any interest to us in analyzing the abortion attitudes of the American electorate. The other 21 observations are either not useful or not of interest – people who may not have answered the question, or their answers were not recorded by the interviewer. Those observations have been coded to **missing** for this variable, meaning that when we analyze this variable, we will not be taking those observations into account. In fact, we probably will want to code the observations in the “other” category to missing, which I will show you how to do below.

Of course, we are interested in the abortion attitudes of the 1,786 people who responded to this survey question only insofar as we can **generalize** from these observations to find something out about the abortion attitudes of the whole American electorate. So, what we really want to know is what percentage of Americans has various positions on the abortion issue. So, far more interesting than the frequencies in the second column are the percentages in the second column. They tell us, for example, that the percentage of Americans who take the pure pro-choice position on abortion (always allow) is far greater than the percentage of Americans who take the pure pro-life position (never allow).

The final column shows the cumulative percentage, which is the percentage of all observations at or below that value of the variable. That may be of some use for variables that have some natural ordering (**ordinal** or **interval** variables), but are not of any use for variables (like religious affiliation or region) that do not have any natural ordering (**nominal** variables). Since the abortion variable is ordered from the most pro-life to the most pro-choice attitude, the cumulative percentage does provide some useful information. For example, it tells us that over 41 percent of Americans have abortion attitudes that typically are considered pro-life (never allow or only allow in the limited circumstances of rape, incest, or danger to the life of the woman).

Adding New Variables to an Existing STATA File

Suppose that after we have downloaded the variables from the 2000 NES data and worked with some of the variables, labeling them and labeling their values, we realize that there are some variables that we want to analyze, but have not included in our data set – for example, attitudes on parental consent for abortion and late-term (or partial birth) abortions. Does that mean that we have to start over and again download all of the relevant variables from the 2000 NES data? No! All we have to do is bring in the new variables using STATA's **merge** command. If, for example, we wanted to add attitudes on parental consent for abortion and late-term (or partial birth) abortions to our nes2000_ps241.dta file, we would do the following:

(1) Go through the steps discussed above to create a new STATA data set including the respondent id and the parental consent (v000702) and late-term abortion (v000705) variables. Let's say you call it nes2000_new.dta. The respondent id must be in both data sets in order to merge them. Merging requires that both data sets have a variable that has a unique value for each observation. The respondent (or case) id is generally the only such variable.

(2) Bring the new data set into STATA and rename the respondent id variable to "caseid."

(3) In order to merge the two data sets on the caseid variable, you have to arrange both data sets so that observations are in the order of the values of the caseid variable. In order to arrange the observations in the new data set this way, use the **sort** command:

sort caseid

Then save the new data set (nes2000_new.dta).

(3) Go into the original data set (nes2000_ps241) and sort that data set by the caseid:

sort caseid

(4) Merge in the new data set using the following command:

merge caseid using "C:\PSCI 241\Fall 2002\nes2000_new"

Note that I had saved the new data set in the following directory: C:\PSCI 241\Fall 2002 on my hard drive. You will need to replace that with the disk drive and directory to which you have saved the new data set. Keep in mind that you will need the quotation marks around the file name for the new data set.

(4) This will create a new variable called _merge. You can run a frequency distribution of _merge in order to see if the two data sets have merge properly. If the two sets of variables have merged properly for each observation, each observation will have a value of 3 on _merge. If everything is ok, you can drop _merge from the data set (see below).

(5) Save the new data set.

Deleting Variables

To delete variables from your data set, simply use the **drop** command, as follows:

```
drop _merge
```

Recoding Variables and Creating New Variables

There are times when we want to recode the values of our variables – we want to reorder the values, we want to eliminate certain values, or we want to combine a large number of values into a smaller number of values. This section gives you an overview of the various scenarios under which you might want to recode your variables and how to do so.

(1) Recoding values to missing

There may be some values of a variable that have not already been coded to missing (not useable) that you want to code to missing. For example, in the abortion attitude variable, you might want to get rid of value number 7 (“other, volunteered”) because it does not have much meaning in terms of the other four values of the variable. To do that, you use the **replace** command to recode variables, and the code for missing values is "."

```
replace abortion=. if abortion==7
```

Note that STATA requires you to use two equal signs the second time that an equal sign appears in a command. We probably want to do the same thing to the view of the Bible variable because it also has a value number 7 for a volunteered “other” response:

```
replace bibview=. if bibview==7
```

(2) Reversing the direction of the variable

There are times when you might want to reverse the direction of your variable so that, for example, it ranges from the most liberal response to the most conservative response rather than from the most conservative response to the most liberal response. Most of the issue variables in the NES range from the most liberal to the most conservative attitude. So, to maintain consistency, we might want to reverse the direction of those variables that range from the most conservative to the most liberal attitude. Abortion attitude is one of those variables. It ranges from the most conservative (pro-life) response to the most liberal (pro-choice) response. To reverse the values of abortion so that higher values represent more conservative responses, you would follow the following steps:

(1) Create a new variable that is equal to the old variable using STATA's **gen** (for generate) command:

gen abortreverse=abortion

(2) Use a series of **replace** commands so that the highest value of the new variable is equal to the lowest value of the old variable, and so forth:

replace abortreverse=1 if abortion==4
replace abortreverse=2 if abortion==3
replace abortreverse=3 if abortion==2
replace abortreverse=4 if abortion==1

(3) Assign new value labels and a variable label to the new variable (that's optional) and ask for a frequency distribution of the new variable:

```
. tab abortreverse

abortion attitude |      Freq.     Percent        Cum.
------------------+-----------------------------------
     always allow |        753       42.83       42.83
other, clear need |        265       15.07       57.91
 rape/incest/life |        525       29.86       87.77
      never allow |        215       12.23      100.00
------------------+-----------------------------------
            Total |       1758      100.00
```

(3) Combining the values of a variable into a smaller number of categories

For some of our variables, we may want to combine the values of the variables into a smaller number of categories. For example, it might be nice to have a party identification variable that has only three categories–Democratic, Independent, Republican–in addition to the 7-category party identification variable we now have. To do that, we would follow these steps:

(a) Ask for a frequency distribution of party identification so we can see what the various values stand for.

```
. tab partyid

                  k1x. party id summary |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
0. strong democrat (1,1,0 in k1, k1a/b, |        346       19.38       19.38
1. weak democrat (1,5/8/9,0 in k1, k1a/ |        274       15.35       34.73
2. independent-democrat (3/4/5/8,0,5 in |        269       15.07       49.80
3. independent-independent (3,0,3/8/9 i |        206       11.54       61.34
4. independent-republican (3/4/5/8,0,1  |        230       12.89       74.23
5. weak republican (2,5/8/9,0 in k1, k1 |        215       12.04       86.27
6. strong republican (2,1,0 in k1, k1a/ |        236       13.22       99.50
7. other. minor party. refuses to say ( |          9        0.50      100.00
----------------------------------------+-----------------------------------
                                  Total |       1785      100.00
```

(b) We probably want to recode value number 7 (other party/minor party/refuses to say) to missing:

replace partyid=. if partyid==7

(c) Create a new variable that will be our new three-category party identification variable

gen partyid3=partyid

(d) Use the replace command to combine the 7 values of “partyid” into 3 values for “partyid3."

replace partyid3=1 if partyid<2
replace partyid3=2 if partyid>1 & partyid<5
replace partyid3=3 if partyid>4 & partyid<7

The first command groups strong and weak Democrats into one category. The second command groups all three types of independents (independents who lean Democratic, independents who lean toward neither party, and independents who lean Republican) into one category. Please note that “<“ means “less than” in STATA, “>” means “greater than,” “&” refers to “and,” and “|” means “or”. The third command groups strong and weak Republicans into one category. Note that I did not just ask STATA to recode all values of “partyid” that are greater than 4 to 3 in “partyid3.” Instead, I asked STATA to recode all values of “partyid” that are greater than 4 AND less than 7 to 3 in “partyid3.” The reason is that STATA assigns missing values “invisible” codes (i.e. we can’t see them) that are usually greater than the largest observed value of the variable (e.g. 9). So, if I simply asked STATA to to recode all values of “partyid” that are greater than 4 to 3 in “partyid3,” STATA would recode both weak and strong Republicans and all missing values to 3 in “partyid3.” So, it is best to set an upper limit when combining the highest values of a variable into a single category (i.e. always say greater than some value AND less than some other value).

(e) (Optional step): Label the new variable and label its values:

label var partyid3 “three-category party ID”

label define partyid3 1 “Democrat” 2 “independent” 3 “Republican”
label values partyid3 partyid3

(f) Ask for a frequency distribution of the new variable:

```
. tab partyid3

three-categ |
  ory party |
         ID |      Freq.     Percent        Cum.
------------+-----------------------------------
   Democrat |        620       34.91       34.91
independent |        705       39.70       74.61
 Republican |        451       25.39      100.00
------------+-----------------------------------
      Total |       1776      100.00
```

We might also want to do something similar with the presidential vote variable, which has the following frequency distribution:

```
. tab presvote

c6. r vote cast for president |      Freq.     Percent        Cum.
------------------------------+-----------------------------------
                  1. al gore  |        590       50.64       50.64
            3. george w. bush |        530       45.49       96.14
             5. pat buchanan  |          3        0.26       96.39
               6. ralph nader |         33        2.83       99.23
           7. other (specify) |          9        0.77      100.00
------------------------------+-----------------------------------
                        Total |       1165      100.00
```

Suppose we wanted to have a variable representing just the two-party presidential vote. We could do the following:

```
. gen presvote2=presvote
(642 missing values generated)

. replace presvote2=0 if presvote==1
(590 real changes made)

. replace presvote2=1 if presvote==3
(530 real changes made)

. replace presvote2=. if presvote>3
(45 real changes made, 45 to missing)

. label var presvote2 "two-party presidential vote"

. label define presvote2 0 "gore" 1 "bush"

. label values presvote2 presvote2

. tab presvote2
```

```
  two-party |
 presidentia |
     l vote |      Freq.     Percent        Cum.
------------+-----------------------------------
       gore |        590       52.68       52.68
       bush |        530       47.32      100.00
------------+-----------------------------------
      Total |       1120      100.00
```

This generates a variable coded 0 for Al Gore voters and 1 for George Bush voters. Supporters of all other candidates have been coded to missing for this variable.

(4) Creating a new variable containing the values of multiple other variables

We still have not created a worship attendance variable because the various categories of worship attendance are included in three separate variables (v000877, v000879, and v000880). Frequency distribution of those three variables yields the following:

```
. tab v000877

 x1. attend |
  religious |
   services |      Freq.     Percent        Cum.
------------+-----------------------------------
     1. yes |       1249       69.62       69.62
     5. no  |        545       30.38      100.00
------------+-----------------------------------
      Total |       1794      100.00

. tab v000879

      x2. attend religious |
        services how often |      Freq.     Percent        Cum.
---------------------------+-----------------------------------
            1. every week  |        479       38.50       38.50
      2. almost every week |        205       16.48       54.98
  3. once or twice a month |        270       21.70       76.69
     4. a few times a year |        282       22.67       99.36
                 5. never  |          8        0.64      100.00
---------------------------+-----------------------------------
                     Total |       1244      100.00

. tab v000880

     x2a. attend relig serv > once/week |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                         1. once a week |        270       56.37       56.37
          2. more often than once a week |        209       43.63      100.00
----------------------------------------+-----------------------------------
                                  Total |        479      100.00
```

So, there are six different values of worship attendance contained in these three variables:

(1) Never attend (5 in v000877 OR 5 in v000879)
(2) Attend a few times a year (4 in v000879)
(3) Attend once or twice a month (3 in v000879)

(4) Attend almost every week (2 in v000879)
(5) Attend once a week (1 in v000880)
(6) Attend more often than once a week (2 in v000880)

To create a worship attendance variable, we would use STATA's **gen** and **replace** commands as follows:

```
. gen attend=1 if v000877==5 | v000879==5
(1254 missing values generated)

. replace attend=2 if v000879==4
(282 real changes made)

. replace attend=3 if v000879==3
(270 real changes made)

. replace attend=4 if v000879==2
(205 real changes made)

. replace attend=5 if v000880==1
(270 real changes made)

. replace attend=6 if v000880==2
(209 real changes made)

. label var attend "worship attendance"

. label define attend 1 "never" 2 "a few times a year" 3 "once or twice a month" 4
"almost every week" 5 "once a week" 6 "more than once a week"

. label values attend attend

. tab attend

   worship attendance |      Freq.     Percent        Cum.
----------------------+-----------------------------------
                never |        553       30.91       30.91
   a few times a year |        282       15.76       46.67
once or twice a month |        270       15.09       61.77
    almost every week |        205       11.46       73.23
          once a week |        270       15.09       88.32
more than once a week |        209       11.68      100.00
----------------------+-----------------------------------
                Total |       1789      100.00
```

We might then want to create a worship attendance variable with fewer categories to make some of our analyses a bit easier. For example, we might want to have three categories: rarely attend (1 and 2 in attend), attend somewhat regularly (3 and 4 in attend), and attend at least once a week (5 and 6 in attend). We would create that variable as follows:

```
. gen attend3=attend
(18 missing values generated)

. replace attend3=1 if attend<3
(282 real changes made)

. replace attend3=2 if attend>2 & attend<5
```

```
(475 real changes made)

. replace attend3=3 if attend>4 & attend<7
(479 real changes made)

. label var attend3 "3-category worship attendance"

. label define attend3 1 "rarely" 2 "somewhat regular" 3 "at least once a week"

. label values attend3 attend3

. tab attend3

  3-category worship |
          attendance |      Freq.     Percent        Cum.
---------------------+-----------------------------------
              rarely |        835       46.67       46.67
    somewhat regular |        475       26.55       73.23
at least once a week |        479       26.77      100.00
---------------------+-----------------------------------
               Total |       1789      100.00
```

Printing and Saving Output

Before we get into statistical analysis in STATA, you should know how to print and save the results of your analysis. You have two options. For either option, you must open a **log** file before you do your analysis.

Option 1: You can print your results directly from STATA:

(1) Before you do your analysis, open the log file: choose the **log** option from the file menu and click on **begin**. STATA will ask you for a name of your log file and you can name it anything you want (e.g. ps241).

(2) Do your analysis. (NOTE: Do not close the log file (as you would if you wanted to save your file and bring it into a word processing program (option 2)) if you want to print it directly from STATA.)

(3) When you are done with your analysis, choose the **view** option from the file menu. A box saying “choose file to view” will open and, if you have opened up a log file, will already have the name of your log file in the “file or url:” line. All you have to do is click on **ok** and STATA will open up a view window containing the contents of your log file (i.e. the results of all of the analyses you have done since you opened the log file).

(4) To print the log file, keep the view window open and choose **print viewer** from the file menu. STATA will open up a print box and you should click on **ok**. STATA will then open up a box called “printer settings” where you can type in headers identifying this analysis that will show up on the printed output. For example, you might type a header of “Analysis for PSCI 241, 3/14/02" so that you can remember when and why you did this analysis when you refer to it later.

However, the headers are just for your convenience. You don't have to type a header. After you have typed a header (or if you have chosen not to type one), click on **ok** and STATA will send your log file to the printer.

Option 2: You can save your results (your log file) to a disk and then open that file in a word processing program.

(1) Before you do your analysis, open the log file: choose the **log** option from the file menu and click on **begin**. STATA will ask you for a name of your log file and you can name it anything you want (e.g. ps241). The difference between this option and option 1 is that you do not want to save the file as STATA's default file type (formatted log). So, before you click "save," go to the "save as type" line and choose "Log (*.log)". This will create a file on your disk with a suffix of ".log" (e.g. ps241.log).

(2) Do your analysis.

(3) When you are done with your analysis, again choose the **log** option from the file menu and click on **close**. You can then open this file (e.g. ps241.log) in a word processor and print it from there.

STATISTICAL ANALYSIS IN STATA

Once we have the variables in our data set up the way we want them, we are ready to begin testing our hypotheses by examining the relationship between our independent and dependent variables. To test our hypotheses, we will use what are known as **sample statistics**. Sample statistics are used to assess the relationship between two variables in a **sample** from a larger population (e.g. the National Election Study interviews a sample of the American electorate) in order to determine whether or not the hypothesis holds true for the entire **population** (here, the American electorate).

There are three things we can do with statistics in order to determine whether or not our hypothesis is correct. The first is to examine the **direction** of the relationship between the independent and dependent variables in our sample. By direction, I mean is the relationship between the two variables a **positive** one (i.e. as one variable increases, the other variable increases) or a **negative** one (i.e. as one variable increases, the other variable decreases)? We have hypothesized a positive relationship between pro-life abortion attitudes and Republican party identification: the more pro-life on abortion attitudes individuals are, the more likely they are to identify with the Republican party. We can use statistics to see if that is true.

The second thing we can do with statistics is to examine the **strength** of the relationship between our independent and dependent variables. Just because the relationship between the independent and dependent variables in the sample (in our case, in the NES data) is in the same direction as the one we hypothesized, that does not necessarily mean that our hypothesis is correct. For example, it may be that individuals with pro-life attitudes are just slightly more likely than individuals with pro-choice attitudes to identify with the Republican party. Such a **weak** relationship between abortion attitudes and party identification in the sample would not support our hypothesis that these two variables are related in the population (in the American electorate). We can use statistics to assess how strong the relationship between two variables is.

The third thing we can do with statistics to test our hypotheses is to assess whether or not we can **generalize** beyond the sample to the entire population of interest. It may be that we observe a strong, positive relationship between abortion attitudes and party identification in the NES sample. However, we are not really interested in the NES sample. We are interested in finding something out about the political attitudes and affiliations of the entire American electorate. So, the next question to answer is can we generalize from what we have found in the NES sample to the entire American electorate?

To answer that question, we turn to what is known as a **test of statistical significance**. Such a statistic tells us how confident we can be that the relationship we observed in the sample holds in the population.

Bivariate Statistics I: Examining the Relationship Between Two Nominal or Ordinal Variables

The statistical techniques used for examining the relationship between only two variables are known as **bivariate statistics**. The easiest way to examine the relationship between two variables is what is known as a **bivariate crosstabulation** or just **crosstab**, which is a table displaying the simultaneous values of two variables. A crosstab tells us the percentage of individuals with each value of one variable that take on the various values of a second variable, and is most appropriate for variables that have a limited number of values. It is not very useful for variables that have a large number of values. That means that it is not appropriate for interval variables or for nominal and ordinal variables that have a large number of categories. It is appropriate for nominal and ordinal variables that have a limited number of categories. For example, it would be far more useful for the three-category party identification variable we created than for the seven-point party identification scale.

To do a crosstab in STATA just use the **tab** command followed by the two variables you want to examine. The following command asks for a crosstab between party identification and abortion attitude.

```
. tab partyid3 abortreverse

three-categ |
  ory party |                 abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |       302         76        156         73 |       607
independent |       308        102        198         75 |       683
 Republican |       132         81        165         63 |       441
------------+--------------------------------------------+----------
```

```
      Total |       742        259        519        211 |      1731
```

As you can see, the values of the first variable you type after **tab** are listed vertically in the lefthand column. The values of the second variable are listed horizontally across the top. As you can also see, if you just type **tab** and the two variables, you just get a frequency count, or the number of observations taking on certain values of both variables. What we would really like to see is the percentage of observations taking on certain values of both variables. To see that, we need to ask STATA for either **row** or **column** percentages. Row percentages are the percentage of each category in the vertical variable (party ID) taking on each value of the horizontal variable (abortion). Column percentages are the percentage of each category in the horizontal variable (abortion) taking on each value of the vertical variable (party ID). It is <u>very</u> important that you be careful to ask for the percentages that you want because the interpretation of column percentages and row percentages is <u>not</u> the same.

For example, let's say that we ask for column percentages:

```
. tab partyid3 abortreverse, col

three-categ |
  ory party |                 abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |       302         76        156         73 |       607
            |     40.70      29.34      30.06      34.60 |     35.07
------------+--------------------------------------------+----------
independent |       308        102        198         75 |       683
            |     41.51      39.38      38.15      35.55 |     39.46
------------+--------------------------------------------+----------
 Republican |       132         81        165         63 |       441
            |     17.79      31.27      31.79      29.86 |     25.48
------------+--------------------------------------------+----------
      Total |       742        259        519        211 |      1731
            |    100.00     100.00     100.00     100.00 |    100.00
```

The first number in each cell is the frequency, the second number is the column percentage. The column percentage is the percentage of people with each abortion attitude that are in each category of party identification. For example, 40.7 percent of people who think that abortion should always be allowed identify with the Democratic party, and 17.79 percent of people who think that abortion should always be allowed identify with the Republican party. Meanwhile, 34.6 percent of people who think that abortion should never be allowed identify with the Democratic party, and 24.9 percent of people who think that abortion should never be allowed identify with the Republican party.

Let's say we ask instead for row percentages:

```
. tab partyid3 abortreverse, row

three-categ |
  ory party |                 abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
```

```
    Democrat |       302         76        156         73 |       607 
             |     49.75      12.52      25.70      12.03 |    100.00 
-------------+--------------------------------------------+----------
 independent |       308        102        198         75 |       683 
             |     45.10      14.93      28.99      10.98 |    100.00 
-------------+--------------------------------------------+----------
  Republican |       132         81        165         63 |       441 
             |     29.93      18.37      37.41      14.29 |    100.00 
-------------+--------------------------------------------+----------
       Total |       742        259        519        211 |      1731 
             |     42.87      14.96      29.98      12.19 |    100.00 
```

The row percentages tell us the percentage of people in each category of party identification who have each attitude on abortion. For example, 49.75 percent of Democrats believe that abortion should always be allowed, while only 29.93 percent of Republicans believe that abortion should always be allowed. Meanwhile, 37.41 percent of Republicans believe that abortion should be allowed only in the cases of rape, incest, or danger to the woman's life, but only 25.7 percent of Democrats have that attitude.

It is also possible to ask for row and column percentages:

```
. tab partyid3 abortreverse, row col

three-categ |
  ory party |              abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
-------------+--------------------------------------------+----------
    Democrat |       302         76        156         73 |       607 
             |     49.75      12.52      25.70      12.03 |    100.00 
             |     40.70      29.34      30.06      34.60 |     35.07 
-------------+--------------------------------------------+----------
 independent |       308        102        198         75 |       683 
             |     45.10      14.93      28.99      10.98 |    100.00 
             |     41.51      39.38      38.15      35.55 |     39.46 
-------------+--------------------------------------------+----------
  Republican |       132         81        165         63 |       441 
             |     29.93      18.37      37.41      14.29 |    100.00 
             |     17.79      31.27      31.79      29.86 |     25.48 
-------------+--------------------------------------------+----------
       Total |       742        259        519        211 |      1731 
             |     42.87      14.96      29.98      12.19 |    100.00 
             |    100.00     100.00     100.00     100.00 |    100.00 
```

The first number in each cell is the frequency, the second number in each cell is the row percentage, and the third number in each cell is the column percentage. That ordering will always be the same regardless of the order in which you type "row" and "col."

However, it is probably a bad idea to ask for both row and column percentages because their interpretation is very different and it is easy to get confused about which is which when you ask for both. A good rule of thumb is to always use column percentages and then determine which variable should be the vertical variable (the first variable in the command) and which variable should be the horizontal variable (the second variable in the command). We usually want the independent variable–the variable we are using to explain changes in the other variable–to be the horizontal variable, and the dependent variable–the variable we are trying to explain with the independent variable–to be the vertical variable. That way, when we ask for column percentages, we get the percentage of each category of the independent variable taking on the various values of the dependent variable. So, the dependent variable should be the first variable in the command, and the independent variable should come second. The rule of thumb can be stated in the following command:

tab DV IV, col

where DV stands for dependent variable and IV stands for independent variable. In our hypotheses, abortion is the independent variable and party identification is the dependent variable. We are hypothesizing that abortion attitude causes people to take on a certain party identification. So, we would want abortion attitude to be the horizontal variable in a crosstab, and party identification to be the vertical variable:

```
. tab partyid3 abortreverse, col

three-categ |
  ory party |                 abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |       302         76        156         73 |       607
            |     40.70      29.34      30.06      34.60 |     35.07
------------+--------------------------------------------+----------
independent |       308        102        198         75 |       683
            |     41.51      39.38      38.15      35.55 |     39.46
------------+--------------------------------------------+----------
 Republican |       132         81        165         63 |       441
            |     17.79      31.27      31.79      29.86 |     25.48
------------+--------------------------------------------+----------
      Total |       742        259        519        211 |      1731
            |    100.00     100.00     100.00     100.00 |    100.00
```

At this point, we can answer the first two questions we need to answer to evaluate our hypotheses: Is the direction of the relationship between the two variables the same as what we hypothesized? And, how strong is the relationship between the two variables. In this case, the direction of the relationship is what we hypothesized: a positive one. The more pro-life an individual's abortion attitudes are, the more likely he/she is to identify with the Republican party. People who think that abortion should be allowed only in the cases of rape, incest, or danger to the life of the woman or who think that abortion should never be allowed are clearly more likely than people who think that abortion should always be allowed to identify with the Republican party and are clearly less likely to identify with the Democratic party.

So, is the relationship between abortion attitudes and party identification a strong one? Well, assessing the strength of relationships in survey data is a bit of an art form. You have to have some experience in doing statistical analyses with survey data to really know what is a strong relationship and what is a weak one. So, this is probably the least important of the three questions that we have to answer to evaluate our hypotheses. But, here I would say that the relationship between abortion attitudes and party identification is of moderate strength. Pro-life people are clearly more Republican than are pro-choice people, but not by a lot. They are only 6 percentage points less likely to be Democrats and only 12 percentage points more likely to be Republicans.

Let’s take a look at some clearer examples of weak and strong relationships. First, take a look at the relationship between gender and abortion attitude:

```
. tab abortion sex, col

                  |    zz1. iwr obs: r
m1/m1.t. abortion |        gender
   self-placement |   1. male  2. female |     Total
------------------+----------------------+----------
      never allow |        78        137 |       215
                  |     10.20      13.80 |     12.23
------------------+----------------------+----------
 rape/incest/life |       242        283 |       525
                  |     31.63      28.50 |     29.86
------------------+----------------------+----------
other, clear need |       126        139 |       265
                  |     16.47      14.00 |     15.07
------------------+----------------------+----------
     always allow |       319        434 |       753
                  |     41.70      43.71 |     42.83
------------------+----------------------+----------
            Total |       765        993 |      1758
                  |    100.00     100.00 |    100.00
```

This is clearly a very weak relationship (if it is a relationship at all). Women are more likely than men to take the most pro-life position (never allow), but only by a very small amount. Women also are more likely than men to take the most pro-choice position, but only by a very small amount. Now, let’s take a look at the relationship between race and party identification:

```
. tab partyid3 race, col

three-categ |  zz2. ftf iwr obs: r
  ory party |         race
         ID |  1. white   2. black |     Total
------------+----------------------+----------
   Democrat |       268         89 |       357
            |     31.60      66.42 |     36.35
------------+----------------------+----------
independent |       366         40 |       406
            |     43.16      29.85 |     41.34
------------+----------------------+----------
 Republican |       214          5 |       219
            |     25.24       3.73 |     22.30
------------+----------------------+----------
      Total |       848        134 |       982
            |    100.00     100.00 |    100.00
```

This is clearly a very strong relationship. Blacks are much more likely than whites to be Democrats, and whites are much more likely than blacks to be Republicans. So, those are examples of very weak and very strong relationships. The relationship between abortion attitudes and party identification is somewhere in between. So, let's say that it is a moderately strong relationship.

We still have not answered the final question that we need to answer about the relationship between abortion attitudes and party identification in order to evaluate our hypothesis: can we generalize from what we have observed in the sample to the entire population? In other words, how confident can we be that there is a relationship between abortion attitudes and party identification in the entire American electorate? In order to answer that question, we turn to a **test of statistical significance**.

There are a lot of different tests of statistical significance all based on various theoretical probability distributions. You can learn about these probability distributions in a statistics class. All we have time for in this class is to learn about their practical applications for testing hypotheses. The test of statistical significance that is used in conjunction with crosstabs is the **chi-square (P^2) test**: a significance test based on the chi-square probability distribution. The chi-square test is based on a comparison of the observed frequencies in a crosstab (the number of observations in each cell of the table–pro-life Democrats, pro-life Republicans, pro-choice Democrats, etc.)) to frequencies that we would expect if there were no relationship between the two variables. In other words, the chi-square test assesses how much different what we observe in the sample is from what we would observe if these two variables were statistically independent.

The chi-square probability distribution then tells us how likely it is that the patterns we observe in the sample would exist if there were in fact, no relationship, in the population. Statistically speaking, for a given value of chi-square, the chi-square distribution indicates the probability (or likelihood) that a P^2 value of at least that magnitude would have been observed if there were no relationship between the two variables. Practically speaking, it indicates the probability (or likelihood) that the two variables are not related in the population. In our example, the chi-square test tells us the probability that abortion attitudes and party identification are not related in the American electorate. If that probability is very low, then we can say that the relationship between income and party ID is **statistically significant** and we can **accept** our hypothesis that the two variables are related in the population. If the probability is high that means that the relationship is **not statistically significant**. In other words, the chances are good that there is no relationship between the two variables in the population and we have to **reject** our hypothesis that they are related.

We can compute a chi-square statistic in STATA simply by adding in **chi2** after the comma in the **tab** command:

```
. tab partyid3 abortreverse, col chi2

three-categ |
  ory party |                 abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |       302         76        156         73 |       607
            |     40.70      29.34      30.06      34.60 |     35.07
------------+--------------------------------------------+----------
independent |       308        102        198         75 |       683
            |     41.51      39.38      38.15      35.55 |     39.46
------------+--------------------------------------------+----------
 Republican |       132         81        165         63 |       441
            |     17.79      31.27      31.79      29.86 |     25.48
------------+--------------------------------------------+----------
      Total |       742        259        519        211 |      1731
            |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(6) =  45.0387   Pr = 0.000
```

STATA gives us what is known as a Pearson's chi-square test. The number in parentheses is called the "degrees of freedom" and it's simply (R-1)(C-1), where R is the number of rows in the crosstab and C is the number of columns in the crosstab. Here, the number of rows is 3 and the number of columns is 4. So, the degrees of freedom are: (3-1)(4-1) = 6. The number 45.0387 is the chi-square statistic. If we were to take a look at a chi-square probability distribution or a table of numbers based on the chi-square distribution, we would see that with 6 degrees of freedom, the probability of observing a chi-square value equal to or greater than 45.0387 if there is no relationship between abortion attitudes and party identification is .000. In other words, the probability (or likelihood) that these two variables are not related in the American electorate is essentially 0.

Fortunately, we don't need a chi-square table to figure out that probability. STATA computes the probability for us and reports it in the form of "**Pr=____**". So, when you run a chi-square test in STATA, you can ignore the chi-square value and degrees of freedom and go directly to the "Pr" value. Here, it tells us that the probability that abortion attitudes and party identification are not related in the American electorate is .000. In other words, there is a 0 percent chance that they are not related.

The probability value is also known as **the level of statistical significance**. Here, it tells us that the relationship is very statistically significant. There is essentially no chance that abortion attitudes and party identification are not related in the American electorate. So, we have positive answers to all of the questions we need to answer to assess our hypotheses: (1) Is the direction of the relationship between abortion attitudes and party identification the same as what we hypothesized? Yes, there is a positive relationship. (2) Is it a strong relationship? Well, it is moderately strong. (3) Is there a statistically significant relationship? In other words, can we be confident in saying that abortion attitudes and party identification are related in the American electorate. Yes. The chi-square test indicates that this is a statistically significant relationship – that the chances that the two variables are not related are essentially zero.

So, what if the probability level of our chi-square test had been higher? What if it had been .01? That means that there is a one percent chance that abortion attitudes and party identification are not related in the populated. Stated differently, it means that if we accept the hypothesis that abortion attitudes and party identification are related in the population, there is a one percent chance that we are wrong. So, are we willing to take that risk of being wrong: a one percent chance? What if the probability is .05? Then, there is a five percent chance that we are wrong. Are we willing to take that risk of being wrong if we accept our hypothesis? Or, do we go ahead and reject the hypothesis because the likelihood of it being wrong is too high? What if the probability is .10? Then, there is a ten percent chance that if we accept our hypothesis, we are wrong. Do we accept the hypothesis or reject it?

Standard practice in statistics is to make the cut-off probability **.05**. In other words, if there is a five percent chance or less that we are wrong if we accept the hypothesis, then we go ahead and accept it. If the chances that we are wrong if we accept our hypothesis are greater than five percent (i.e. the probability is greater than .05), then we reject the hypothesis. So, in the language of statistics, when the probability that two variables are <u>not</u> related in the population is **.05 or less**, we say there is a **statistically-significant relationship** between the two variables. If the probability that the two variables are not related in the population is **greater than .05**, we say that there is **not a statistically-significant relationship** between the two variables.

Let's take a look at another example. Suppose we hypothesized that southerners are more likely than people who live outside of the South to identify with the Republican party. If we compute a crosstab of the relationship between southern residence (a variable I generate based on the region variable in the 2000 NES) and party identification and ask for a chi-square test of the statistical significance of that relationship, we get the following:

```
. tab partyid3 south, col chi2
```

```
three-categ |
  ory party |   region of residence
         ID | non-south      south |     Total
------------+----------------------+----------
   Democrat |       388        232 |       620
            |     34.22      36.14 |     34.91
------------+----------------------+----------
independent |       462        243 |       705
            |     40.74      37.85 |     39.70
------------+----------------------+----------
 Republican |       284        167 |       451
            |     25.04      26.01 |     25.39
------------+----------------------+----------
      Total |      1134        642 |      1776
            |    100.00     100.00 |    100.00

          Pearson chi2(2) =   1.4478   Pr = 0.485
```

The chi-square test tells us that the probability that there is no relationship between southern residence and party identification is .485. In other words, there is a 48.5 percent chance that there is no relationship between southern residence and party identification in the American electorate. This chance is way too high for us to accept the hypothesis that the two variables are related. We have to reject our hypothesis.

Can you think of why southern residence and party identification are not related? We might suspect that they would be given that the South has become much more Republican in recent decades and has become solidly Republican in presidential elections. One reason that they may not be related for the whole electorate is that the South has the highest percentage of African-Americans of any region in the country, and African-Americans identify overwhelmingly with the Democratic party. So, it might be true that southern whites are more likely than white people in other regions to identify with the Republican party, but when we include all races in our analysis, the large number of African-Americans in the South pulls the region in a Democratic direction. To see if that is true, we might want to conduct our analysis only for whites (only when our “race” variable equals 1). We can do that by adding an **if statement** to our command:

```
. tab partyid3 south if race==1, col chi2

three-categ |
  ory party |   region of residence
         ID | non-south      south |     Total
------------+----------------------+----------
   Democrat |       199         69 |       268 
            |     33.11      27.94 |     31.60 
------------+----------------------+----------
independent |       262        104 |       366 
            |     43.59      42.11 |     43.16 
------------+----------------------+----------
 Republican |       140         74 |       214 
            |     23.29      29.96 |     25.24 
------------+----------------------+----------
      Total |       601        247 |       848 
            |    100.00     100.00 |    100.00 

          Pearson chi2(2) =   4.6555   Pr = 0.098
```

Now, we observe a relationship more like the one we would expect. Southern whites are less likely than non-southern whites to identify with the Democratic party, and are more likely than non-southern whites to identify with the Republican party. However, we still would not accept the hypothesis that southerners are more likely than non-southerners to identify with the Republican party. The relationship is a fairly weak one–southern whites are only about six percentage points less Democratic and more Republican than are non-southern whites. And, the chi-square test does not meet standard levels of statistical significance. The probability that there is no relationship between southern residence and party identification among white Americans is .098. There is about a 10 percent chance that these two variables are not related.

One word of caution in doing crosstabs and chi-square tests is that the tendency people have is to go right to the chi-square and not look at the crosstab. The problem with that is that the chi-square statistic only tests the hypothesis that the two variables are related. It does not tell you anything about the direction of that relationship. For example, the chi-square statistic tells you the likelihood that abortion attitudes and party identification are related in the population. It tells you nothing about whether pro-life individuals are more Republican or more Democratic than people who hold pro-choice views on abortion. To ascertain the direction of the relationship between the two variables, we have to look at the crosstab.

In other words, just the fact that the probability value for the chi-square test is very low doesn't necessarily mean that we should accept our hypothesis that the more pro-life individuals' abortion attitudes are, the more likely they are to identify with the Republican party. If the probability value was very low and pro-life people are more likely than pro-choice people to identify with the Democratic party, we would have to reject our hypothesis.

The following table sums up whether you should accept or reject your hypotheses based on cross-tabulations and chi-square tests of significance.

Hypothesis Tests Using Crosstabs and Chi-Square

Direction of relationship	Probability	Accept or reject?
Same as hypothesis	<.05	Accept
Opposite of hypothesis	<.05	Reject
Same or opposite of hypothesis	>.05	Reject

Bivariate Statistics II: Examining the Relationship Between Two Ordinal or Interval Variables

Sometimes a crosstab is not very helpful in assessing the direction, strength, and statistical significance of the relationship between two variables. That is particularly the case for interval-level variables, or other variables that have a large number of categories. For example, suppose we wanted to examine the relationship between abortion attitudes and the feeling thermometer ratings of George W. Bush in the 2000 NES. Feeling thermometer ratings range from 0 to 100 and can take on any whole number value in between. So, they potentially have 101 different values – far too many for a crosstab to provide any meaningful evidence. When variables such as these are involved in our hypotheses, we need to use other statistical methods to test those hypotheses.

One such statistic is a **correlation coefficient**. A correlation coefficient assesses the extent to which there is a **linear** or **straight-line** relationship between two variables (see pp. 377 and 378 in the Johnson, Joslyn, and Reynolds reading). In other words, a correlation coefficient tells us the extent to which increases in one variable are associated with increases or decreases in another variable. For our example, a correlation tells us the extent to which increases in pro-life attitudes on abortion are associated with increases or decreases in positive feelings about George W. Bush. Because the correlation coefficient indicates the extent to which increases or decreases in one variable are associated with increases or decreases in another variable, it is only appropriate to examine relationships between variables that have some meaningful ordering (for which increases and decreases have some meaning): ordinal and interval variables, but not nominal variables.

The correlation coefficient **ranges from -1 to 1**. If the correlation coefficient is close to zero, that means that there is no relationship between the two variables. For example, the correlation between education and thermometer ratings of Bush in the 2000 NES is -.01. That means that increases in education are associated with little or no change in ratings of George W. Bush.

As the correlation coefficient gets closer to 1, that means that there is a strong positive relationship between the two variables – increases in one variable are associated with increases in the other variable. For example, the correlation between the seven-point party identification scale (ranging from strong Democrat to strong Republican) and feeling thermometer ratings of Bush is .55. That means that increases in identification with the Republican party are associated with large increases in positive feelings toward Bush. (Although a correlation coefficient of .55

is a long way from 1, it is still very strong in survey data. See below for more on this point.)

As the correlation coefficient gets closer to -1, that means that there is a strong negative relationship between the two variables – increases in one variable are associated with decreases in the other variable. For example, the correlation between party identification (ranging from strong Democrat to strong Republican) and feeling thermometer ratings of Al Gore is -.61. That means that increases in identification with the Republican party are associated with large decreases in positive feelings toward Gore.

One thing to note is that correlation coefficients in survey data rarely approach either -1 or 1. The reason is that there is a good deal of measurement error in survey data. Variables in survey data like the ideology and party identification variables or the abortion attitude variable are, for various reasons (vague or inexact question wording, top of the head answers by respondents, coding mistakes, etc.) often not truly accurate reflections of individuals' true attitudes or orientations. That measurement error decreases the extent to which we can explain changes in one variable through changes in another variable, and increases the extent to which changes in particular variables are random (not systematically related to identifiable factors). So, a correlation of .55 between party identification and Bush ratings in survey data suggests a very strong positive relationship. A correlation of -.61 between party identification and Gore ratings in survey data suggests a very strong negative relationship.

So, there are not any real formal guidelines for what is a strong correlation and what is not, but based on what political scientists have found with NES survey data, I would say the following general guidelines apply.

Correlation	Strength of relationship
.5 to 1, -.5 to -1	Very strong
.35 to .5, -.35 to -.5	Strong
.25 to .35, -.25 to -.35	Modest
.10 to .25, -.10 to .25	Weak
0 to .10, 0 to -.10	None

We can compute correlation coefficients in STATA by just using the **corr** command. For example, correlations between party identification and Bush ratings and between party identification and Gore ratings are as follows:

```
. corr partyid bushft
(obs=1736)

             |  partyid   bushft
-------------+------------------
     partyid |   1.0000
      bushft |   0.5563   1.0000


. corr partyid goreft
(obs=1748)

             |  partyid   goreft
-------------+------------------
     partyid |   1.0000
      goreft |  -0.6065   1.0000
```

You can ignore the values of 1.0000 in the table. They simply tell us that the correlation between a variable and itself is 1. The coefficients we are interested in are the ones in the bottom left-hand corner. These are the correlations between party identification and feeling thermometer ratings of the 2000 presidential candidates.

Correlation coefficients are what are known in statistics as **measures of association**. They tell us the direction and strength of the relationship between two variables in the sample. But, they do not tell us anything about statistical significance. They do not, for example, tell us the chances that the relationship between party identification and Bush ratings does not exist in the American electorate.

We can compute the level of statistical significance of a correlation coefficient in STATA by using the **pwcorr** command with ",sig" at the end.

```
. pwcorr partyid bushft, sig

             |  partyid   bushft
-------------+------------------
     partyid |   1.0000
             |
             |
      bushft |   0.5563   1.0000
             |   0.0000
             |
```

The correlation coefficient is again in the bottom left-hand corner, and the level of statistical significance is directly below it. Here, the level of statistical significance of the relationship between party identification and Bush ratings is .0000. That tells us that there is a 0 percent chance that there is no relationship between these two variables in the American electorate. It is a very statistically significant relationship.

What about the relationship between education and feelings toward Bush?

```
. pwcorr education bushft, sig
```

```
             | educat~n   bushft
-------------+------------------
   education |   1.0000 
             |
             |
      bushft |  -0.0137   1.0000 
             |   0.5648
             |
```

Well, the very weak correlation coefficient suggests that there is no relationship between these two variables, and the level of statistical significance of .5648 confirms that. The probability that education levels and Bush ratings are not related in the American electorate is .5648. This relationship clearly is not statistically significant.

Bivariate Statistics III: Examining the Relationship Between a Nominal Variable and an Interval Variable by Creating "Dummy" Variables

What if we want to examine the relationship between a nominal variable and an interval variable? Correlation coefficients are not appropriate for variables that have no natural ordering, so what do we do? There are several alternatives, but one is to create several **dichotomous** or **dummy** variables out of a nominal variable. A dummy variable is a variable that is equal to one for the presence of some trait and is equal to zero for the absence of that trait. For example, the dummy variable ("south") that I created for southern residence is equal to one for residents of the South and zero for everyone else. The gender variable in the NES is already a dummy variable (although we might want to recode it to equal 0 for men and 1 for women rather than 1 for men and 2 for women) because it has only two categories.

Dummy variables can be used in correlation analyses because they do have some natural ordering. As we move from zero to one, we are moving from the absence of the trait to the presence of that trait. So, if we want to examine the relationship between a nominal variable and an interval variable, we can create several dummy variables out of the nominal variable, and examine the correlation between those dummy variables and the interval variable.

For example, suppose we hypothesized that there is a relationship between region and ratings of Bush: southerners are more likely than people who live in other regions to support Bush and people who live in the Northeast are more likely than people who live in other regions to oppose Bush. Because the region variable is a nominal variable, we cannot correlate it with the Bush thermometer ratings. However, we can create dummy variables for southern residence and northeastern residence and correlate those with the Bush ratings:

```
. pwcorr south bushft, sig

             |    south   bushft
-------------+------------------
       south |   1.0000 
             |
             |
      bushft |   0.1250   1.0000 
             |   0.0000
             |

. pwcorr northeast bushft, sig

             | northe~t   bushft
-------------+------------------
   northeast |   1.0000 
             |
             |
      bushft |  -0.0421   1.0000 
             |   0.0775
             |
```

This tells us that there is, as we hypothesized, a positive relationship between living in the South and positive evaluations of George W. Bush. The correlation is fairly weak. But, it is positive and statistically significant. There is a 0 percent chance that there is no relationship between southern residence and support for Bush. On the other hand, we have to reject our hypothesis that living in the northeast is related to negative evaluations of Bush. The correlation between northeastern residence and Bush ratings is negative. But, it is very weak and not statistically significant. There is a 7.75 percent chance that there is no relationship in the American electorate between living in the northeast and having negative feelings toward George W. Bush. This chance is too high for us to accept our hypothesis.

Multivariate Statistics I: Controlling for One Other Variable in a Crosstab

So far, we have examined the bivariate (two-variable) relationship between abortion attitudes and party identification, and that analysis has told us that there is a positive relationship between pro-life abortion attitudes and Republican party identification, that the relationship is moderately strong, and that the relationship is statistically significant. Does that mean that we can accept our hypothesis that the more pro-life an individual's abortion attitudes are, the more likely he/she is to identify with the Republican party? The answer is "no" because we have not yet controlled for other variables that may affect the relationship between abortion attitudes and party identification. As discussed above, some factors, such as region or religious beliefs, may cause both abortion attitudes and party identification so that the observed relationship between those two variables is really a spurious one. Other factors, such as liberal-conservative ideology, may intervene between abortion attitudes and party identification so that the relationship between those two variables is an indirect, rather than a direct, one. Statistical analyses of the relationships between multiple variables (i.e. more than two variables) are known as **multivariate statistical analyses**.

To determine whether or not there really is a relationship between abortion attitudes and party identification (i.e. the relationship is not spurious) and whether that relationship is a direct or indirect one, we need to **control** for the other variables that may be related to both abortion attitudes and party identification. In other words, we need to **hold** these variables **constant** (hold them at the same values) so that an observed relationship between changes in abortion attitudes and changes in party identification can not be due to changes in these other variables. In the case of a crosstab between two variables, the way that we control for a third variable is simply by examining the crosstab between the independent and dependent variables within each category of the third variable.

Let's start with the possibility that the relationship between abortion attitudes and party identification is spurious because both are caused by religious beliefs. To assess that possibility, the first thing we would want to do is make sure that both abortion attitudes and party identification are related to religious beliefs (as measured by the view of the Bible question in the NES):

```
. tab abortreverse bibview, col chi2

                  | s5/s5.t. bible is word of god or
                  |               men
abortion attitude | 1. the bi  2. the bi  3. the bi |     Total
------------------+---------------------------------+----------
     always allow |       126        393        183 |       702 
                  |     21.32      46.40      75.62 |     41.79 
------------------+---------------------------------+----------
other, clear need |        73        157         26 |       256 
                  |     12.35      18.54      10.74 |     15.24 
------------------+---------------------------------+----------
 rape/incest/life |       241        240         28 |       509 
                  |     40.78      28.34      11.57 |     30.30 
------------------+---------------------------------+----------
      never allow |       151         57          5 |       213 
                  |     25.55       6.73       2.07 |     12.68 
------------------+---------------------------------+----------
            Total |       591        847        242 |      1680 
                  |    100.00     100.00     100.00 |    100.00 

          Pearson chi2(6) = 315.2187   Pr = 0.000

. tab partyid3 bibview, col chi2

three-categ | s5/s5.t. bible is word of god or
  ory party |               men
         ID | 1. the bi  2. the bi  3. the bi |     Total
------------+---------------------------------+----------
   Democrat |       228        273         95 |       596 
            |     38.71      31.86      39.26 |     35.31 
------------+---------------------------------+----------
independent |       203        342        116 |       661 
            |     34.47      39.91      47.93 |     39.16 
------------+---------------------------------+----------
 Republican |       158        242         31 |       431 
            |     26.83      28.24      12.81 |     25.53 
------------+---------------------------------+----------
      Total |       589        857        242 |      1688 
            |    100.00     100.00     100.00 |    100.00 

          Pearson chi2(4) =  32.2705   Pr = 0.000
```

There clearly is a very strong and statistically significant relationship between view of the Bible and abortion attitudes. People who believe that the Bible is the literal Word of God are much less likely to say that abortion should always be allowed and much more likely to say that abortion should never be allowed than are people who believe that the Bible is not the Word of God. The probability that these two variables are not related in the American electorate is basically zero. The relationship between view of the Bible and party identification is not as strong. People who see the Bible as the literal Word of God are actually no less likely to identify with the Democratic party than are people who do not think that the Bible is the Word of God. However, there is a relationship between the two variables. Biblical literalists are clearly more likely than people who do not see the Bible as the Word of God to identify themselves as Republicans, and the relationship between view of the Bible and party identification is very statistically significant.

So, both abortion attitudes and party identification are related to religious beliefs. That means that it is possible that the relationship between abortion attitudes and party identification is spurious because both are caused by religious beliefs. To see if that is true, we need to examine the relationship between abortion attitudes and party identification while controlling for view of the Bible. To do that in STATA, we first **sort** the data by the control variable, and then run our crosstab **by** the control variable.

```
. sort bibview

. by bibview: tab partyid3 abortreverse, col chi2

________________________________________________________________________________
-> bibview = 1. the b

three-categ |
  ory party |              abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |        54         30         85         55 |       224
            |     43.55      41.67      35.86      37.41 |     38.62
------------+--------------------------------------------+----------
independent |        49         25         75         51 |       200
            |     39.52      34.72      31.65      34.69 |     34.48
------------+--------------------------------------------+----------
 Republican |        21         17         77         41 |       156
            |     16.94      23.61      32.49      27.89 |     26.90
------------+--------------------------------------------+----------
      Total |       124         72        237        147 |       580
            |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(6) =  10.6151   Pr = 0.101

________________________________________________________________________________
```

```
-> bibview = 2. the b

three-categ |
  ory party |                abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |       157         38         57         15 |       267 
            |     40.46      24.68      23.95      26.32 |     31.90 
------------+--------------------------------------------+----------
independent |       150         62        100         21 |       333 
            |     38.66      40.26      42.02      36.84 |     39.78 
------------+--------------------------------------------+----------
 Republican |        81         54         81         21 |       237 
            |     20.88      35.06      34.03      36.84 |     28.32 
------------+--------------------------------------------+----------
      Total |       388        154        238         57 |       837 
            |    100.00     100.00     100.00     100.00 |    100.00 

          Pearson chi2(6) =  31.5406   Pr = 0.000

_______________________________________________________________________________
-> bibview = 3. the b

three-categ |
  ory party |                abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |        78          7          8          2 |        95 
            |     43.33      29.17      28.57      40.00 |     40.08 
------------+--------------------------------------------+----------
independent |        83         10         16          3 |       112 
            |     46.11      41.67      57.14      60.00 |     47.26 
------------+--------------------------------------------+----------
 Republican |        19          7          4          0 |        30 
            |     10.56      29.17      14.29       0.00 |     12.66 
------------+--------------------------------------------+----------
      Total |       180         24         28          5 |       237 
            |    100.00     100.00     100.00     100.00 |    100.00 

          Pearson chi2(6) =   9.5605   Pr = 0.144

_______________________________________________________________________________
-> bibview = .

three-categ |
  ory party |                abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |        13          1          6          1 |        21 
            |     26.00      11.11      37.50      50.00 |     27.27 
------------+--------------------------------------------+----------
independent |        26          5          7          0 |        38 
            |     52.00      55.56      43.75       0.00 |     49.35 
------------+--------------------------------------------+----------
 Republican |        11          3          3          1 |        18 
            |     22.00      33.33      18.75      50.00 |     23.38 
------------+--------------------------------------------+----------
      Total |        50          9         16          2 |        77 
            |    100.00     100.00     100.00     100.00 |    100.00 

          Pearson chi2(6) =   4.2892   Pr = 0.638
```

This analysis shows that the relationship between abortion attitudes and party identification is statistically significant only for individuals who believe that the Bible is the Word of God, but should not be taken literally. In this group, individuals who say that abortion should always be allowed are much more likely to be Democrats and much less likely to be Republicans than are individuals who say that abortion should never be allowed. The probability that there is not a statistically significant relationship between abortion attitudes and party identification among all members of the electorate who see the Bible as the Word of God, but not literally true is essentially zero.

However, the relationship between abortion attitudes and party identification is not statistically significant for individuals who see the Bible as the literal Word of God or for individuals who believe that the Bible is not the Word of God. In the first group, individuals who say that abortion should never be allowed are less likely to be Democrats and more likely to be Republicans than are individuals who say that abortion should always be allowed. The relationship approaches standard levels of statistical significance, but there is a 10 percent chance that there is no relationship between abortion attitudes and party identification among all citizens who view the Bible as literally true. That chance is too high for us to conclude with any confidence that these two variables are related for all Biblical literalists.

Among people who believe that the Bible is not the Word of God, there is very little variation in abortion attitudes. The large majority (180 out of a total of 237) has the most pro-choice view on abortion. Since there is little variation in abortion attitudes within this group, it is not surprising that variation (or change) in abortion attitudes does not have a statistically significant relationship with variation in party identification. There is a 14.4 percent chance that there is no relationship between these two variables among all citizens who do not see the Bible as the Word of God.

The final crosstab is for bibview=. Remember that "." in STATA refers to a missing value. So, this crosstab is for individuals who have a missing value (a value that is not meaningful) on the view of the Bible variable. You should ignore this last crosstab.

Based on these results, should we conclude that abortion attitudes and party identification are related or that the relationship between the two variables is spurious because both are caused by religious beliefs? Well, the evidence is mixed. In fact, a problem with this method of controlling for a third variable is that it often does not provide a definitive answer about whether or not the relationship between an independent variable and a dependent variable is statistically significant even when controlling for another variable. I will discuss another, better method of controlling for other variables below. But, for now, we simply have to conclude that the relationship that we observed between abortion attitudes and party identification when we did not control for any other variables is due <u>in part</u> to the two variables' mutual relationship with religious beliefs. However, the observed relationship is <u>not due entirely</u> to the two variables' mutual relationship with religious beliefs because that relationship remains statistically significant for certain types of religious beliefs.

Now, let’s examine the possibility that liberal-conservative ideology intervenes between abortion attitudes and party identification, that abortion attitudes affect party identification indirectly through ideology. To do that, I first created a new variable “ideology3" in which the seven-point ideology scale is collapsed into three categories: liberals (1-3 on the scale), moderates (4 on the scale), and conservatives (5-7 on the scale). We can now examine the relationship between abortion attitudes and party identification within each category of this ideology variable.

```
. sort ideology3

. by ideology3: tab partyid3 abortreverse, col chi2

_______________________________________________________________________________
-> ideology3 = liberal

three-categ |
  ory party |             abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |        67          9         14          7 |        97
            |     62.62      32.14      46.67      77.78 |     55.75
------------+--------------------------------------------+----------
independent |        34         15         16          0 |        65
            |     31.78      53.57      53.33       0.00 |     37.36
------------+--------------------------------------------+----------
 Republican |         6          4          0          2 |        12
            |      5.61      14.29       0.00      22.22 |      6.90
------------+--------------------------------------------+----------
      Total |       107         28         30          9 |       174
            |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(6) =  20.8150   Pr = 0.002

_______________________________________________________________________________
-> ideology3 = moderate

three-categ |
  ory party |             abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |        28         16         21          5 |        70
            |     29.47      34.04      33.87      62.50 |     33.02
------------+--------------------------------------------+----------
independent |        51         20         33          2 |       106
            |     53.68      42.55      53.23      25.00 |     50.00
------------+--------------------------------------------+----------
 Republican |        16         11          8          1 |        36
            |     16.84      23.40      12.90      12.50 |     16.98
------------+--------------------------------------------+----------
      Total |        95         47         62          8 |       212
            |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(6) =   6.2489   Pr = 0.396
```

```
_______________________________________________________________________________
-> ideology3 = conservative

three-categ |
  ory party |                abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |        15          5         19          5 |        44 
            |     17.86      11.36      18.63      14.71 |     16.67 
------------+--------------------------------------------+----------
independent |        29         10         31         14 |        84 
            |     34.52      22.73      30.39      41.18 |     31.82 
------------+--------------------------------------------+----------
 Republican |        40         29         52         15 |       136 
            |     47.62      65.91      50.98      44.12 |     51.52 
------------+--------------------------------------------+----------
      Total |        84         44        102         34 |       264 
            |    100.00     100.00     100.00     100.00 |    100.00 

          Pearson chi2(6) =   5.8487   Pr = 0.440

_______________________________________________________________________________
-> ideology3 = .

three-categ |
  ory party |                abortion attitude
         ID | always al  other, cl  rape/ince  never all |     Total
------------+--------------------------------------------+----------
   Democrat |       192         46        102         56 |       396 
            |     42.11      32.86      31.38      35.00 |     36.63 
------------+--------------------------------------------+----------
independent |       194         57        118         59 |       428 
            |     42.54      40.71      36.31      36.88 |     39.59 
------------+--------------------------------------------+----------
 Republican |        70         37        105         45 |       257 
            |     15.35      26.43      32.31      28.13 |     23.77 
------------+--------------------------------------------+----------
      Total |       456        140        325        160 |      1081 
            |    100.00     100.00     100.00     100.00 |    100.00 

          Pearson chi2(6) =  34.3171   Pr = 0.000
```

This analysis suggests that ideology does play an intervening role between abortion attitudes and party identification. In other words, it does provide at least a partial explanation for the relationship between abortion attitudes and party identification. The relationship is statistically significant among liberals. Liberals with pro-life attitudes are more likely to identify as Republicans and less likely to identify as Democrats than are liberals with pro-choice attitudes on abortion, and there is only a .2 percent chance that abortion attitudes and party identification are not related among all American liberals. However, the relationship between abortion attitudes and party identification is not statistically significant for either moderates or conservatives.

We can also control for a third variable in examining the correlation between two ordered variables (i.e. interval or ordinal variables). For example, if we ask STATA to compute the correlation between abortion attitudes and feeling thermometer ratings of Bush, we get the following:

```
. pwcorr abortreverse bushft, sig

             | abortr~e   bushft
-------------+------------------
abortreverse |   1.0000 
             |
             |
      bushft |   0.2752   1.0000 
             |   0.0000
             |
```

So, there is a positive relationship between pro-life abortion attitudes and positive evaluations of Bush, and the relationship is very statistically significant. However, there is a possibility that the relationship is spurious because both variables may be caused by religious beliefs (view of the Bible). To see if that is true, we can compute correlations between abortion attitudes and Bush ratings for each different view of the Bible.

```
. sort bibview

. by bibview: pwcorr abortreverse bushft, sig

______________________________________________________________________________
-> bibview = 1. the b

             | abortr~e   bushft
-------------+------------------
abortreverse |   1.0000 
             |
             |
      bushft |   0.1539   1.0000 
             |   0.0002
             |

______________________________________________________________________________
-> bibview = 2. the b

             | abortr~e   bushft
-------------+------------------
abortreverse |   1.0000 
             |
             |
      bushft |   0.2630   1.0000 
             |   0.0000
             |

______________________________________________________________________________
-> bibview = 3. the b

             | abortr~e   bushft
-------------+------------------
abortreverse |   1.0000 
             |
             |
      bushft |   0.1858   1.0000 
             |   0.0041
             |

______________________________________________________________________________
-> bibview = .
```

```
             | abortr~e   bushft
-------------+------------------
abortreverse |   1.0000
             |
             |
      bushft |   0.1922   1.0000
             |   0.1008
             |
```

The result is that for each of the meaningful views of the Bible (ignoring view of the Bible equals "missing"), there is a positive and statistically significant correlation between abortion attitudes and party identification. In other words, even when we hold view of the Bible constant, increases in pro-life abortion attitudes are still related to increases in positive evaluations of Bush. So, we can conclude that the relationship between abortion attitudes and view of the Bible is not spurious because both variables are related to view of the Bible.

Multivariate Statistics II: Controlling for Multiple Variables in Multiple Regression Analysis

So far, we have controlled for the effects of other variables on the relationship between our independent and dependent variables through what Johnson, Joslyn, and Reynolds (p. 396) call **control by grouping**. We group the observations according to their values on the third variable and then observe the original relationship within each of these groups. For example, we observed the relationship between abortion attitudes and party identification within each of the categories of the view of the Bible variable.

There are two major problems with this form of statistical control. One problem is that it is really only feasible to control for one variable at a time. If we control for multiple variables (e.g. view of the Bible and ideology), we have far too many groupings (e.g. liberals with a literal view of the Bible, liberals who think the Bible is the Word of God but is not literally true, liberals who think that the Bible is not the Word of God, conservatives with a liberal view of the Bible, conservatives who think the Bible is the Word of God but is not literally true...........) to make any sense of whether the control variable really affects the relationship between the independent and dependent variables. And, the number of observations in each of these various groupings is often too small for us to be able to generalize from our results to the entire population.

The second problem is one that I already have noted. Control by grouping often gives us mixed evidence for the effect that a third variable has on the relationship between our independent and dependent variables. For example, we found that the relationship between abortion attitudes and party identification was statistically significant for people in the middle category of view of the Bible, but not for people in the first and third categories. This does not provide us with a clear picture of whether abortion attitudes and party identification really are related or whether their observed relationship is due to their mutual relationship with view of the

Bible.

A more desirable method of statistical control is known as **multiple regression analysis**, which is appropriate only when our dependent variable has some natural ordering and works better when the dependent variable has a large number of values (e.g. if party identification was your dependent variable, you would want to use the seven-category variable rather than the three-category variable). Multiple regression analysis indicates how much a one-unit change (e.g. moving from 1 to 2 on the NES abortion scale or from 5 to 6 on the NES ideology scale) in an independent variable changes the dependent variable **when all other variables in the model have been held constant**. The controlling is done by mathematical manipulation, not by literally grouping subjects together. **Control by adjustment** is a form of statistical control in which a mathematical adjustment is made to assess the impact of a third variable.

For example, we may identify a number of factors that may affect the relationship between abortion attitudes and party identification: view of the Bible, ideology, worship attendance, age, education, and income. Rather than computing crosstabs or correlations between abortion attitude and party identification while controlling for each of these other variables separately, we can simultaneously control for all of these variables by conducting a multiple regression analysis in which party identification is the dependent variable and abortion attitude, view of the Bible, ideology, worship attendance, age, education, and income are the independent variables. In the language of statistics, we **regress** party identification on abortion attitude and all of the control variables. The **regression coefficient** on abortion attitude will indicate how much a one-unit change in abortion attitude (e.g. moving from always allow abortion to allow in the case of a clear need or moving from allow in the cases of rape, incest, or danger to the woman's life to never allow) changes party identification when all of the control variables are held constant. To conduct a regression analysis in STATA, simply use the **reg** (for regress) command, followed immediately by the name of the dependent variable, then by all of the independent variables.

```
. reg partyid abortreverse bibview ideology educ income age

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  6,   519) =   33.23
       Model |  639.620387     6  106.603398           Prob > F      =  0.0000
    Residual |  1665.10395   519  3.20829277           R-squared     =  0.2775
-------------+------------------------------           Adj R-squared =  0.2692
       Total |  2304.72433   525  4.38995111           Root MSE      =  1.7912

------------------------------------------------------------------------------
     partyid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
abortreverse |   .1632979   .0881513     1.85   0.065    -.0098793    .3364752
     bibview |   .2138517   .1390113     1.54   0.125    -.0592424    .4869457
    ideology |   .7127354   .0602737    11.82   0.000     .5943249    .8311458
   education |   .1866806   .0574137     3.25   0.001     .0738889    .2994723
      income |   .0411101   .0211935     1.94   0.053    -.0005255    .0827457
         age |  -.0116869   .0049611    -2.36   0.019    -.0214333   -.0019406
       _cons |  -1.618772   .5434986    -2.98   0.003      -2.6865   -.5510446
------------------------------------------------------------------------------
```

The first column of the table (under "coef.") provides the **regression coefficient** for each independent variable. This tells us that pro-life abortion attitudes do have a positive relationship to Republican party identification. An increase of one on the abortion scale is associated with an increase of .16 on the seven-point party identification scale, when all of the other independent variables are controlled or held constant. An increase of one on the seven-point ideology scale (ranging from strong liberal to strong conservative) is associated with an increase of .71 on the party identification scale, when all of the other independent variables are held constant. Age, on the other hand, has a negative relationship with party identification. An increase of one year in age is associated with a decrease of .01 on the party identification scale.

One thing to note here is that after the last independent variable, age, there is a row labeled _cons. This refers to the **regression constant** in our model (see p. 406 in Johnson, Joslyn, and Reynolds). It simply tells us the value of the dependent variable when all of the independent variables in our regression analysis are equal to zero. Because it is often impossible for all of the independent variables to be equal to zero (e.g. in our regression analysis, many of the variables (income, age, abortion attitude) do not have any values equal to zero), we will often get a constant that is outside of the actual range of the dependent variable. That is the case here. Party identification ranges from 0 to 6, but the regression constant is -1.62. So, it is usually best to pay no attention to the constant.

The third column (under "t") provides what is known as a **t-statistic**. A t-statistic is a test of statistical significance much like the chi-square statistic. The only difference is that they are based on different theoretical probability distributions. The chi-square statistic is based on the chi-square distribution. The t-statistic is based on something called the **Student's t** distribution. What is important for both statistics is the **level of statistical significance** associated with the test of statistical significance. For the t-statistic in a multiple regression analysis, this is provided in the fourth column (under "P>|t|"). Statistically speaking, for a given value of t, the t distribution indicates the probability (or likelihood) that a t value of at least that magnitude would have been observed if changes in the independent variable were associated with no changes in the dependent variable (i.e. the two variables are not related). . Practically speaking, it indicates the probability (or likelihood) that the independent variable has no effect on the dependent variable in the population.

For example, the level of statistical significance of the effect of abortion attitudes on party identification, holding all of the other independent variables constant, is .065. This indicates that the probability that abortion attitude has no effect on party identification (i.e. changes in abortion attitudes are associated with no change in party identification) in the American electorate is .065. There is a 6.5 percent chance that we are wrong if we conclude that abortion attitudes do affect party identification. This is larger than the standard cut-off probability of .05, but it is not far off. The probability that ideology has no effect on party identification, holding all of the other independent variables constant, is essentially zero. Ideology clearly has a statistically significant effect on party identification. The probability that age has no effect on party identification, holding all of the other independent variables constant, is .019. There is a 1.9 percent chance that age does not effect party identification in the American electorate. This is also a statistically significant relationship.

Much as with crosstabs and the chi-square test of statistical significance, you should be careful not to accept or reject your hypotheses based on only on significance levels in multiple regression. Suppose we hypothesized that the older individuals are, the more likely they are to identify with the Republican party. The effect of age on party identification is clearly statistically significant. So, do we accept our hypothesis? No. We have to reject it because the direction of the relationship is not the one we hypothesized. It is negative, not positive. Increases in age are associated with decreases in identification with the Republican party.

One problem with what we have done so far is that we have no way of comparing the **size of the effects** of the various independent variables on the dependent variable, party identification. All of the independent variables have different scales, so that a one-unit increase in one variable means something entirely different than a one-unit increase in another variable. For example, the abortion variable only ranges from 1 to 4, whereas age ranges from 18 to 97. So, it appears that the effect of abortion attitude on party identification (a regression coefficient of .16) is larger than the effect of age on party identification (a regression coefficient of -.01). However, that may not be true. If we move from the lowest value (1 for most pro-choice) of abortion to the highest value (4 for most pro-life) of abortion, the increase in Republican party identification is .48 (4-1=3; 3 x .16 = .48). If we move from the lowest age (18) to the highest age (97), the decrease in Republican party identification is .79 (97-18=79; 79 x -.01 = -.79). So, it is impossible to compare the size of the effects of various independent variables on a dependent variable with ordinary regression coefficients.

To compare the size of the effects of various independent variables on a dependent variable in multiple regression analysis, we need to compute **standardized regression coefficients**. Standardized regression coefficients put all of the variables on the same scale so that we can compare the relative importance of each independent variable in explaining change in the dependent variable (see p. 408 in Johnson, Joslyn, and Reynolds). To compute a standardized regression coefficient in STATA, simply place a comma after the last independent variable, and after the comma, type **beta** (standardized regression coefficients are sometimes called “beta weights”).

```
. reg partyid abortreverse bibview ideology educ income age, beta

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  6,   519) =   33.23
       Model |  639.620387     6  106.603398           Prob > F      =  0.0000
    Residual |  1665.10395   519  3.20829277           R-squared     =  0.2775
-------------+------------------------------           Adj R-squared =  0.2692
       Total |  2304.72433   525  4.38995111           Root MSE      =  1.7912

------------------------------------------------------------------------------
     partyid |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
abortreverse |   .1632979   .0881513     1.85   0.065                  .080945
     bibview |   .2138517   .1390113     1.54   0.125                 .0686128
    ideology |   .7127354   .0602737    11.82   0.000                 .4872282
   education |   .1866806   .0574137     3.25   0.001                 .1378397
      income |   .0411101   .0211935     1.94   0.053                 .0790478
         age |  -.0116869   .0049611    -2.36   0.019                -.0892315
       _cons |  -1.618772   .5434986    -2.98   0.003                        .
------------------------------------------------------------------------------
```

The standardized regression coefficients are in the last column of the table (under “Beta”). They tell us that ideology has a much stronger effect on party identification than any of the other independent variables in our model. The next strongest effect is that of education. The size of the effects of all of the other variables on party identification is about the same.

Multivariate Statistics III: Multiple Regression Models with Dichotomous (Dummy) Independent Variables

Much like correlation coefficients, multiple regression analysis requires that all of the independent variables and the dependent variable have some natural ordering (interval and ordinal variables). So, what if we want to include a nominal variable in a multiple regression analysis. What we have to do is create a series of **dichotomous** or **dummy** variables for the various categories of the nominal variable, just like we did in correlation analysis. The only difference here is that we need to create **N-1 dummy variables** where N is the number of categories of the nominal variable.

For example, if we wanted to include region in our multiple regression analysis, we would create 3 dummy variables because region has 4 categories. So, let’s say that we create dummy variables for the South (a variable coded 1 for southern residents and 0 for residents of all other regions), the Northeast (a variable coded 1 for northeastern residents and 0 for residents of all other regions), and the West (a variable coded 1 for western residents and 0 for residents of all other regions). The only region for which we do not create a dummy variable is the Midwest (NES calls it "north central"). When we include the three dummy variables for regions in our multiple regression analysis, the coefficients on the dummy variables compare the level of the dependent variable (party identification) for that region to the level of the dependent variable for the excluded region (the Midwest – the region for which we did not include a dummy variable in the regression analysis). **It is <u>important</u> to remember that regression coefficients on dummy variables are not interpreted in the same way as regression coefficients on ordinal or interval variables. They always indicate the difference in the level of the dependent variable between the particular category of a nominal variable represented by the dummy variable and the category of that nominal variable that is not included in the regression analysis.**

You also should remember that a variable such as "sex" is also a dummy variable. There are two categories of gender: male and female. So, in essence what we have is a dummy variable for one of those categories: female (a variable coded 1 for female and 0 for male). So, the coefficient on sex is the difference in the level of the dependent variable between women and men.

So, let's say we want to include region and gender as independent variables in our regression analysis. We would simply include dummy variables for all but one of the categories of the nominal variable in our regression analysis. So, we include three dummy variables for region and the one dummy variable for gender.

```
. reg partyid abortreverse bibview ideology educ income age south northeast west sex,
beta

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F( 10,   515) =   20.35
       Model |  652.764821    10  65.2764821           Prob > F      =  0.0000
    Residual |  1651.95951   515  3.20768838           R-squared     =  0.2832
-------------+------------------------------           Adj R-squared =  0.2693
       Total |  2304.72433   525  4.38995111           Root MSE      =   1.791

------------------------------------------------------------------------------
     partyid |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
abortreverse |   .1589062   .0890851     1.78   0.075                 .0787681
     bibview |   .2370152   .1421531     1.67   0.096                 .0760447
    ideology |   .7071991   .0608434    11.62   0.000                 .4834436
   education |   .1873416   .0574683     3.26   0.001                 .1383278
      income |   .0414719   .0214233     1.94   0.053                 .0797435
         age |  -.0112552   .0049888    -2.26   0.024                -.0859352
       south |   .0429753   .2123489     0.20   0.840                 .0098274
   northeast |  -.3856661   .2459159    -1.57   0.117                -.0717371
        west |  -.0725735   .2348457    -0.31   0.757                -.0145914
         sex |  -.0717966   .1621485    -0.44   0.658                -.0171254
       _cons |   -1.54493   .6152828    -2.51   0.012                        .
------------------------------------------------------------------------------
```

The results show that none of the coefficients on our dummy variables are statistically significant. So, the coefficient on the dummy variable for the South is positive, but not statistically significant. If it were statistically significant, we would say that the difference in party identification between residents of the South and residents of the Midwest (the excluded category of region)–holding abortion attitude, view of the Bible, ideology, education, income, age, and gender constant–is .04. However, that difference is not statistically significant. The probability that there is not a difference between the party identifications of southern residents and midwestern residents in the whole American electorate is .84. So, we have to conclude that there is no difference in the party identifications of southern residents and midwestern residents, holding all of the other independent variables constant. The same thing applies to the differences between residents of the Northeast and residents of the Midwest and to the differences between residents of the West and residents of the Midwest.

The effect of gender on party identification, when we hold all of the other independent variables constant, is not statistically significant. If the effect were statistically significant, the negative coefficient on sex would indicate that, holding all of the other independent variables constant, women have party identifications that are .07 points lower on the party identification scale than are those of men. However, the probability that there is not a difference between the party identifications of men and women, when we hold all of these other independent variables constant, in the American electorate is .658. So, we have to conclude that there is no difference in the party identifications of women and men.

Multivariate Statistics III: Multiple Regression Models with Dichotomous (Dummy) Dependent Variables

When the dependent variable in our analysis is a dichotomous (or dummy) variable (e.g. the two-party presidential vote), the most appropriate analysis is not multiple regression analysis, but **logistic regression** or **logit** for short. I think this method is a bit too mathematically involved for the short time we have in this class. So, if you are interested in it, refer to pp. 412-430 in the Johnson, Joslyn, and Reynolds reading. Otherwise, simply use multiple regression analysis.

For example, if we wanted to examine the relationship between abortion attitude and the two-party presidential vote (coded 0 for Gore and 1 for Bush) in 2000 while controlling for a range of other variables, we would regress the presidential vote on abortion attitude and all of these other independent variables. The coefficient on abortion attitude would indicate **the change in the likelihood (or probability) of voting for George Bush for a one-unit change in abortion attitude, holding all of the other independent variables constant**.

```
. reg presvote2 abortreverse bibview ideology educ income age sex south, beta

      Source |       SS       df       MS              Number of obs =     362
-------------+------------------------------           F(  8,   353) =   20.05
       Model |   28.232254     8  3.52903175           Prob > F      =  0.0000
    Residual |  62.1323869   353  .176012428           R-squared     =  0.3124
-------------+------------------------------           Adj R-squared =  0.2968
       Total |  90.3646409   361  .250317565           Root MSE      =  .41954

------------------------------------------------------------------------------
   presvote2 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
abortreverse |   .0498309   .0252735     1.97   0.049                 .1034029
     bibview |  -.0490402   .0413409    -1.19   0.236                -.0637441
    ideology |   .1648388   .0172167     9.57   0.000                 .4723796
   education |  -.0001614   .0166487    -0.01   0.992                -.0004825
      income |   .0103892   .0059551     1.74   0.082                 .0841197
         age |  -.0010689   .0014186    -0.75   0.452                -.0339363
         sex |   .0082463   .0466316     0.18   0.860                 .0082463
       south |    .055588   .0477586     1.16   0.245                 .0522642
       _cons |  -.2882226   .1783946    -1.62   0.107                        .
------------------------------------------------------------------------------
```

The results indicate that the effect of abortion attitude on the likelihood of voting for Bush, holding view of the Bible, ideology, education, income, age, sex, and southern residence constant, is positive and statistically significant. An increase of one unit in pro-life attitudes is associated with an increase of .05 in the probability of voting for Bush. The probability that abortion attitude did not have an effect on voting for Bush in the American electorate is .049 – i.e. a 4.9 percent chance that abortion attitude had no effect on the probability of voting for Bush.