POLI 300 ***Handouts #10*** Fall 2006

ASSOCIATION BETWEEN VARIABLES: CROSSTABULATIONS

Suppose we want to do research on the following bivariate hypothesis: *the more interested people are in politics, the more likely they are to vote* (sentence #13 in Problem Sets #3A and #9). In the manner of Handout #9, we can diagram this as follows:

LEVEL OF POLITICAL INTEREST + WHETHER/NOT VOTED [*inds*]
(Low or High) =========> (Yes or No)

The dependent variable is intrinsically *dichotomous* (two-valued). Suppose we also use a very imprecise measure for the independent variable that is also dichotomous (with just "Low" vs. "High" values.) *Note*: recall that, given a dichotomous variable like WHETHER/NOT VOTED with "yes" and "no" values, the "no" value is conventionally deemed to be "low" and "yes" to be "high," which allows us to characterize this hypothesized association as positive.

We design an NES type of survey with n = 1000 respondents and collect data on both variables. As a first step we do univariate analysis on each variable — in particular, we construct these two univariate absolute frequency tables:

LEVEL OF POLITICAL INTEREST		WHETHER/NOT VOTED	
Low	500	No	500
High	500	Yes	500
Total	1000	*Total*	1000

The first and very important point is that these two *univariate frequency distributions* provide *no evidence whatsoever* bearing on the *bivariate hypothesis* of interest. It is possible that every respondent with a "low" value on INTEREST fails to vote and that every respondent with a "high" value on INTEREST does vote (which would powerfully confirm our hypothesis). But, as a logical possibility, the reverse could also be true — that is, it might be that every respondent with a "low" value on INTEREST does vote and that every respondent with a "high" value on INTEREST fails to vote (which would totally contradict our hypothesis). And of course there is a huge range of intermediate possibilities.

Crosstabulations and Association Between Variables

We can analyze the relationship or association between two *discrete* variables such as these by means of a *crosstabulation* (also called a *contingency table*) — what might be called a *joint* (or *bivariate*) *frequency table* as it is in effect *two intersecting frequency tables*. Recall that in a regular (univariate) frequency distribution (Handout #5), the rows of the table correspond to the values of the variable (usually with an additional row at the bottom that shows totals). In a crosstabulation, the rows of the table correspond to the values of one variable that is naturally called the *row variable* (again usually with an additional row at the bottom that shows column totals). But the table is likewise divided into a number of columns corresponding to the values of the other variable that is naturally called the *column variable* (sometimes with one additional column at the right that shows row totals). Each (interior) *cell* of the table is defined by the *intersection* of a row and column and corresponds to a particular *combination of values*, one for each variable. As with a univariate frequency table, the most basic piece of information associated with each cell is the corresponding *absolute frequency* — that is, the number of cases that have that particular combination of values on the two variables.

The general template for the crosstabulation between the dichotomous variables WHETHER OR NOT VOTED by LEVEL OF INTEREST is shown in Table 1 below.

TABLE 1. CROSSTABULATION OF WHETHER OR NOT VOTED BY LEVEL OF INTEREST

LEVEL OF INTEREST

		Low	High	Row Total
VOTED?	No	No & Low	No & High	Total No
	Yes	Yes & Low	Yes & High	Total Yes
	Col. Total	Total Low	Total High	Grand Total

It is conventional to make the *independent* variable the *column* variable and the *dependent* variable the *row* variable, as we have done here. The darker shaded portions of are not part of the table itself but simply show the value labels for each variable. The lighter shaded portions of the table show the row and column totals, which are simply the univariate frequencies of each variable taken by itself; they are often called the *marginal frequencies*, because they appear on the "margins" (edges) of the table. The unshaded cells in the *interior* of the table constitute the *2* × *2* crosstabulation proper. It is the *joint frequency distribution over the cells in this interior of the table that tell us whether and how the two variables are related or associated*. We can infer little (in general) or nothing (in this case, because of its "uniform marginals" — see below) about the interior of the crosstabulation from its marginal frequencies alone. All sorts of different patterns are compatible with these univariate frequencies, as we can see in Tables 1A-1F that follow.

	Low	High	Total
No	???	???	500
Yes	???	???	500
Total	500	500	1000

Table 1A. $a = ?$

	Low	High	Total
No	500	0	500
Yes	0	500	500
Total	500	500	1000

Table 1B. $a = +1$

	Low	High	Total
No	350	150	500
Yes	150	350	500
Total	500	500	1000

Table 1C $a \approx +.5$

	Low	High	Total
No	250	250	500
Yes	250	250	500
Total	500	500	1000

Table 1D $a = 0$

	Low	High	Total
No	150	350	500
Yes	350	150	500
Total	500	500	1000

Table 1E. $a \approx -.5$

	Low	High	Total
No	0	500	500
Yes	500	0	500
Total	500	500	1000

Table 1F. $a = -1$

Table 1A shows the generic table. The cell entries are unspecified and can be filled in any way that is consistent with the marginal frequencies. Table 1B displays a *perfect positive association* between the two variables so, for any measure of association a, we have $a = +1$. Table 1C displays a *weak positive association* between the two variables so a equals something like +0.5 — in any case, some positive value intermediate between 0 and +1. Table 1D displays the *absence of any association* between the two variables, so $a = 0$. Table 1E displays a *weak negative association* between the two variables, so $a \approx -0.5$. Table 1F displays a *perfect negative association* between the two variables, so we have $a = -1$.

If the natural ordering of the values of a variable runs from Low to High, the entirely standard (and sensible) convention is that Low to High on the column variables runs from left to right; the less standard (and less sensible) convention is that Low to High on the row variable runs from top to bottom (so the Low-Low "origin" of the table is its "northwest corner"). More generally, if a crosstabulation pertains to variables with matching values, the convention is that these values are listed in a common ascending or descending order from left to right for the column variable and from top to bottom for the row variable. Given this convention, a positive association between the two variables means that the joint frequencies are concentrated (highly if the positive association strong, less so is the positive association is weaker) in the cells along the so-called *main diagonal* of the table running from the "northwest" corner (No & Low in Table 1) to the "southeast" corner (Yes & High in Table 1), as is illustrated in panels 1A and 1B. A negative association between the two variables means the joint frequencies are concentrated in the cells along the *off- diagonal* of the table running from the "southwest" corner (No & High in Table 1) to the "northeast" corner (Yes & Low in Table 1), as is illustrated in panels 1E and 1F. If there is little or no association between the variables, the joint frequencies will be more or less uniformly dispersed among all cells in the table, as is illustrated by panel 1C.

Table 1 provides the simplest possible example of a crosstabution, in a number of respects.

First, it is a *2×2 table* with just two rows and two columns, because both variables are dichotomous. Many tables have more than two rows and/or columns, because they crosstabulate variables with more than two possible values. Second, Table 1 is *square*, with the same number of rows and columns, but tables may have an unequal number of rows and columns (in which case the "diagonals" are a bit less clearly defined). Third, Table 1 has *uniform marginal frequencies*, i.e., the same number of cases (500) in each row and in each column. Obviously real data is messier than this.

Constructing Crosstabulations

We now consider how actually to construct a crosstabulation from raw data, continuing to focus on the same hypothesis that relates political interest and the likelihood of voting. The Student Survey includes somewhat relevant data, namely responses to a question (V9 on a recent survey) that asks students about their level of interest in the upcoming (or most recent) Presidential election and another (V7 on a recent survey) asking whether or not they voted in the most recent Presidential election. One potential problem is that a lot of data on the latter question may be missing, because quite a few students were not eligible to vote at the time. But our immediate purpose is simply to demonstrate how to construct a crosstabulation from scratch, so we proceed with these two variables

First we need to set up a crosstabulation *template* or *worksheet* for this pair of variables, as is shown below. We create a row for each value of the row variable and a column for each value of the column variable. We also need a row and column for potential missing data, and we should add another row and column for the marginal frequencies, which can be entered in advance if we know both univariate frequencies already (as in the previous hypothetical example). We should always be careful to label the variables and their values, and it is helpful to the reader to give the name a name in this manner: DEPENDENT VARIABLE By INDEPENDENT VARIABLE. Thus our worksheet looks like this (ignore the cell entries for the moment):

TABLE 2A: WHETHER OR NOT VOTED BY LEVEL OF INTEREST

LEVEL OF INTEREST (V9)

VOTED? (V7)	Not much 1	Somewhat 2	Very Much 3	NA 9	***Row Total***
No, not eligible 1		#16	#22		5
Yes 2		#2 #3 #6	#1 #4 #5 #7 #8		33
No (though eligible) 3	#4				3
DK 4			#25		1
NA 9			#15		1
Column Total	3	13	27	0	43

The next step is to process the raw Student Survey data [from an earlier semester], not on a univariate basis for V9 and V7 separately, but *on a bivariate basis for V9 and V7 jointly*. To do this we look at the V9 and V7 columns *simultaneously* and, for each case, note its combination of coded values for V6 and V9 respectively. Looking at the first eight cases, the data from an earlier semester was as follows:

Case ID	*V6*	*V9*
1	3	2
2	2	2
3	2	2
4	1	3
5	3	2
6	2	2
7	3	2
8	3	2

We put a tally mark (or better — and for the same reasons identified in Handout #5 for univariate frequency distributions — put the case ID #) in the appropriate cell of the table. (The first eight cases have been so tallied in Table 2A above, plus a few other cases exhibiting different combinations of values.) Once we have processed all the cases, we convert the tally marks or count of ID#s into absolute frequencies. This produces the following crosstabulation:

TABLE 2B:WHETHER OR NOT VOTED BY LEVEL OF INTEREST

LEVEL OF INTEREST (V9)

VOTED? (V10)	Not much 1	Somewhat 2	Very Much 3	NA 9	***Row Total***
No, not eligible 1	0	3	2	0	5
Yes 2	1	9	23	0	33
No (though eligible) 3	2	1	0	0	3
DK 4	0	0	1	0	1
NA 9	0	0	1	0	1
Column Total	3	13	27	0	43

Next we should remove the missing data row and column, since data that is missing on one or other or both variables can tell us nothing about the association between them. The same applies to the "effectively missing data" that appears in rows 1 and 4. (Respondents in these rows answered

Question 9 but they gave answers that do not bear on the hypothesis of interest, i.e., they either didn't remember whether they voted [row 4] or were not eligible to vote [row 1], regardless of their level of interest.) Next we interchange the "Yes" and "No" rows to match the format of Table 1. Finally, let's *recode* LEVEL OF INTEREST to make it dichotomous (in the manner of Table 1) by combining columns 1 and 2 into a single "Low" value and labeling column 3 "High." The result of these adjustments is that we have a version of Table 2 that is set up in the manner of Table 1. (However, the marginals are far from uniform.) Note that we have removed the value codes and the non-descriptive variable names (i.e., V9 and V7) and have deleted irrelevant rows and columns, so the format is identical to that of Table 1.

TABLE 2C. WHETHER OR NOT VOTED BY LEVEL OF INTEREST

LEVEL OF INTEREST

VOTED?	Low	High	***Total***
No	3	0	3
Yes	10	23	33
Total	13	23	36

a ≈+0.15

Source: POLI 300 Student Survey, Fall 2006

I used SPSS to compute a number of measures of association, such as are discussed in Weisberg, Chapter 12. Most of them have positive but very weak values. In general, the actual association between the variables in the Student Survey data is somewhere between the hypothetical Table 1C and 1D above (though of course the number of cases is *much* smaller). But the main problem we have in using this Student Survey data to assess the hypothesis is that, in this data, both univariate frequencies (especially VOTED?) are highly skewed (rather than being uniform), which tends to mask any association between the variables.

Let us work one more example using Student Survey data. Consider sentence #14 from Problem Sets #3A and #9, which can be stated formally as

DIRECTION OF IDEOLOGY ==========> DIRECTION OF VOTE
(Liberal to Conservative) (Dem. vs. Rep.)

The Student Survey includes appropriate data to test this hypothesis. V27 (Question 27) provides a standard measure of DIRECTION OF IDEOLOGY. Measuring DIRECTION OF VOTE is a bit more problematic, but we can use V8 (Question 8), noting that it refers to *preference*, not to an actual *vote*, in the recent 2004 Presidential election. Code values 4 and 5 must be excluded as missing data, and we will exclude code value 3 (Nader) also, since the hypothesis above codes DIRECTION OF VOTE simply as DEM vs. REP. Now we set up a 2 × 5 table with PRESIDENTIAL PREFERENCE as the row (dependent) variable and IDEOLOGY as the column (independent) variable, and process the Student Survey data in a manner parallel to the previous example. Since IDEOLOGY values run from left to right to left, let's rearrange the rows representing the values of PRESIDENTIAL PREFERENCE into the same "left" (top) to "right" (bottom) ordering. Once we do this, we expect to see a (so to speak) "positive" association between the two variables, i.e., the more conservative a student's ideology, the more conservative his or her vote. Here is the resulting crosstabulation. (Remember, student respondents who gave an "Other" or "DK" responses on V10 are excluded as effectively missing.)

TABLE 3. PRESIDENTIAL PREFERENCE BY IDEOLOGY

IDEOLOGY

PRES PREF	Liberal	Slightly Liberal	Moderate	Slightly Cons.	Conser- vative	***Total***
Kerry	8	11	6	1	0	26
Bush	0	0	1	3	7	11
Total	8	11	7	4	7	37

$a \approx 0.7$

Source: POLI 300, Student Survey, Fall 2006

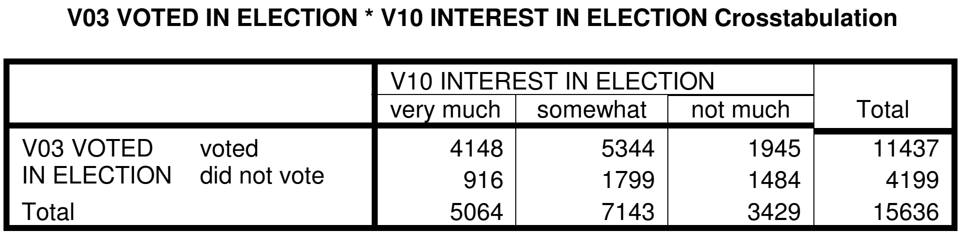
Measures of association calculated by SPSS mostly range from about 0.6 to 0.9, so overall the association between variables is about 0.7, i.e., a much stronger association than in Table 2C. In a 2 × 5 table like this, the notion of a (main) diagonal is rather murky, but notice that, when we look at how cases are distributed over columns, the "center of gravity" moves consistently downwards as we scan the table from left to right

SPSS Crosstabulations

As you would expect, SPSS can construct crosstabulations very readily. Instructions are set out in the Handout on *Using Setups 1972-2004 NES Data and SPSS for Windows* and SPSS tables are illustrated in the accompanying handout on *Data Analysis Using SETUPS and SPSS*.

First, we present the SPSS crosstabulation of SETUPS/NES data (with all nine election years pooled together) for the variables that best measure LEVEL OF INTEREST and WHETHER/NOT VOTED and thus is parallel to Table 2C for Student Survey data.

TABLE 4



V03 VOTED IN ELECTION * V10 INTEREST IN ELECTION Crosstabulation

		V10 INTEREST IN ELECTION			Total
		very much	somewhat	not much	
V03 VOTED IN ELECTION	voted	4148	5344	1945	11437
	did not vote	916	1799	1484	4199
Total		5064	7143	3429	15636

Note that SPSS arranges the rows and columns according to the numerical codes for the values of the variables.[1] Most measures of association for this table are quite low — on the order of a ≈ + 0.2. This is because the distribution of cases with respect to the dependent (row) variable is so lopsided. (Even among the "not much interested" respondents, a substantial majority of claim to have voted.)

TABLE 5

V04 PRESIDENTIAL VOTE * V34 R'S OWN IDEOLOGY Crosstabulation

		V34 R'S OWN IDEOLOGY					Total
		liberal	slightly liberal	moderate	slightly conserv	conserv	
V04 PRESIDENTIAL VOTE	Dem	971	794	1315	465	307	3852
	Rep	115	271	1292	1222	1719	4619
Total		1086	1065	2607	1687	2026	8471

Here I have excluded voters for "Other" Presidential candidates, since over the 1972-2004 period such candidates constitute an ideologically mixed bag. Measures of association range from about + 0.6 to + 0.8, generally similar to the student data.

[1] However, when I ran this table, it became evident that the value labels for V10 in the data file are listed in the reverse order [appearing to reverse the direction of the association]. On checking this further, I determined that this error goes back to the original SETUPS/NES 1970-1992 data that I updated. I have corrected this error in the data file on my computer (and thus in Table 4) but it may not be correct in the data file that the student PCs access.