Revised July 2018

Crosstabulations of Two Categorical Variables

This set of notes shows how to use Stata to crosstabulate two categorical variables. It assumes that you have set Stata up on your computer (see the "Getting Started with Stata" handout), and that you have read in the set of data that you want to analyze (see the "Reading in Stata Format (.dta) Data Files" handout).

In Stata, most tasks can be performed either by issuing commands within the "Stata command" window, **or** by using the menus. These notes illustrate both approaches, using the data file "GSS2016.DTA" (this data file is posted here: https://canvas.harvard.edu/courses/53958).

A crosstabulation is a two-variable frequency distribution, showing the number of times each pair of values occurs together. One may enhance a crosstabulation in numerous ways, through calculation of percentages/relative frequencies, measures of association telling how strongly the variables are linked to one another, and inferential tests for the presence of an association (in particular, the chi-square test for statistical independence). All of these enhancements can be added via Stata.

To obtain a crosstabulation using the "Stata Command" window, issue the following command:

```
tab <rowvar> <colvar>,<options>
```

where you fill in the names of the variables you want to study in place of "<rowvar>" and "<colvar>", and add appropriate keywords in place of "<options>" to choose the supplemental calculations you want to see. The variable you designate as "<rowvar>" will appear in the rows of the crosstabulation and the one you designate as "<colvar>" will be in the columns.

For example, the following command crosstabulates the variable "postlife**"** (belief in life after death, rows) against the variable "sex" (columns), with options asking Stata to calculate percentages within columns of the table ("column" option).

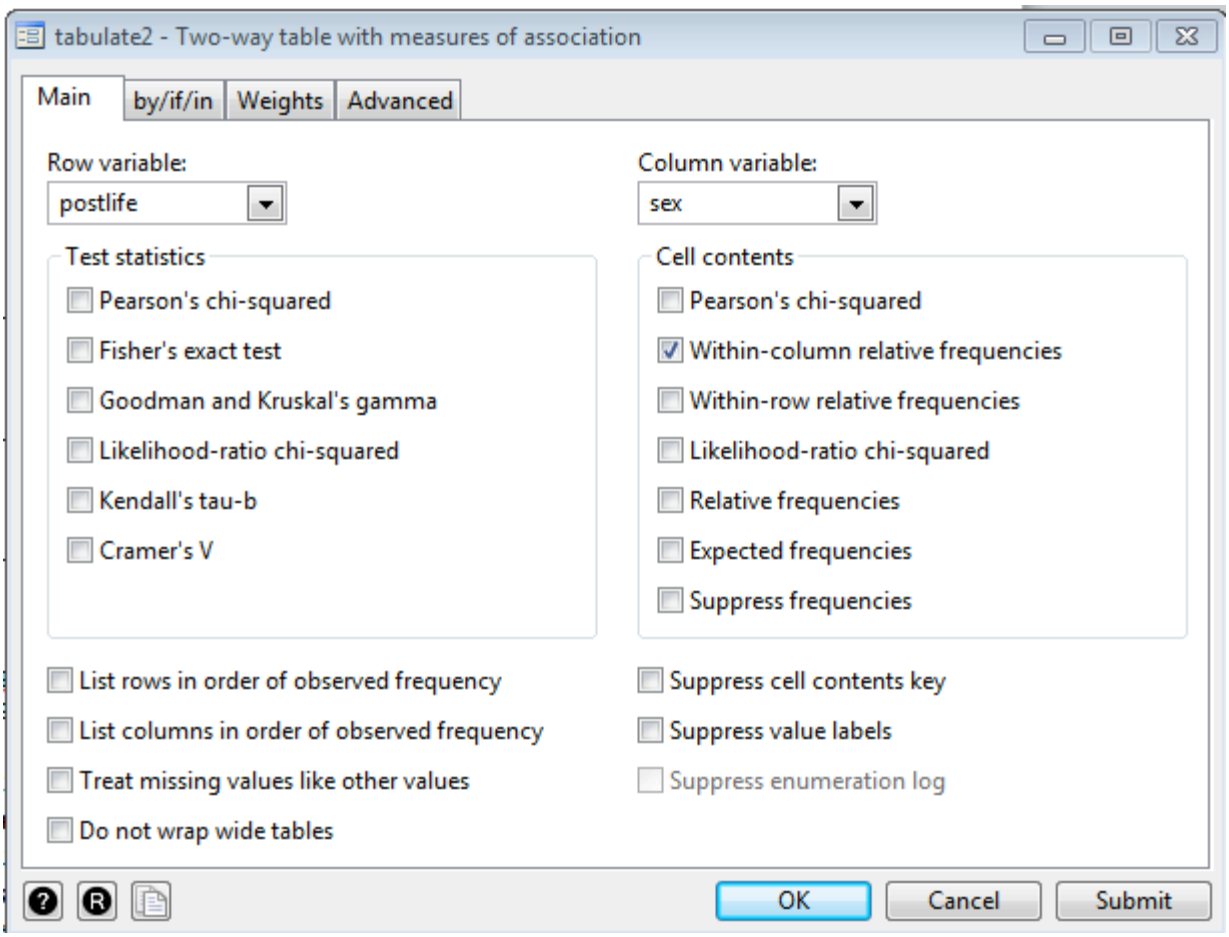
```
tab postlife sex,column
```

This is enough to describe the association in the data set, if you have your "explanatory" variable in columns and your "response" variable in rows. If the explanatory variable is in the rows and the response in the columns, then "row" percentages are usually a better choice.

Using the Stata menus, you produce a crosstabulation as follows:

click on "Statistics"
click on "Summaries, tables, & tests"
click on "Frequency Tables"
click on "Two-way tables with measures of association"

A window like the one shown on the next page will open up:



Fill in the name of the variable you want to put in the rows of your table in the "Row variable:" box, and the name of the one you want in the columns in the "Column variable:" box. Then check boxes to select the optional calculations you want. In this example, percentages (relative frequencies) within columns is selected.

After you have chosen the options you want, click "OK."

Either way (via the command line or the menu), the output shown on the next page appears in the "Stata Results" window:

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

 belief in |
life after |    respondents sex
     death |      male     female |     Total
-----------+----------------------+----------
       yes |       859      1,230 |     2,089
           |     75.09      85.18 |     80.72
-----------+----------------------+----------
        no |       285        214 |       499
           |     24.91      14.82 |     19.28
-----------+----------------------+----------
     Total |     1,144      1,444 |     2,588
           |    100.00     100.00 |    100.00
```

Comparing the percentages within the first row of this table shows that women are about 10 percentage points more likely than are men to believe in life after death.

The "Key" box at the upper left tells the contents of each cell in the crosstabulation; here the top number is the observed cell frequency, and the bottom number is the relative frequency or percentage, calculated within columns of the table. For example, in the "male, yes" cell, there are 859 respondents, who constitute 75.09% of the 1144 male respondents in the study.

If, after some experience, you decide you don't need to see the key each time you run a crosstabulation, you can suppress it by checking the "Suppress the cell contents key" box on the screen shown on the prior page, or–equivalently–by adding the "nokey" option to the options list on the command line.

Some enhancements

Moving beyond describing the association in the data set, we may want to assess its statistical significance using the Pearson chi-square statistic for testing the hypothesis of statistical independence ("chi2" option), or describe its magnitude via a "measure of association" such as gamma. By adding options to the command line, or checking them within the menu shown above, this is easily accomplished.

Here is a more elaborate analysis of a crosstabulation, without the key this time. The command for it is

```
tab postlife sex, column chi gamma nokey
```

with results as shown on the next page:

```
belief in |
life after |    respondents sex
     death |      male     female |     Total
-----------+----------------------+----------
       yes |       859      1,230 |     2,089
           |     75.09      85.18 |     80.72
-----------+----------------------+----------
        no |       285        214 |       499
           |     24.91      14.82 |     19.28
-----------+----------------------+----------
     Total |     1,144      1,444 |     2,588
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =  41.7761   Pr = 0.000
                    gamma =  -0.3120  ASE = 0.045
```

The chi-square statistic has a low p level, indicating that we can be quite confident in asserting that these sample results reflect an association in the population, not just the sample. The value of gamma tells us that there is a moderate negative association between being female and not believing in life after death (this awkward language is due to the ways in which the "polarities" were assigned to the variables in this example; a less awkward phrasing is that there is a moderate negative association between being male and believing in life after death).