Revised July 2018

Confidence Intervals About a Mean or Proportion

This set of notes tells how to use Stata to produce confidence intervals for a population proportion for a dichotomous variable, or the population mean of a continuous variable. 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 confidence interval gives you two numbers that define a range of values that (with high probability) contains the value of a parameter in "the population," based on one's observation of a statistic (here the sample proportion or mean) for a random sample selected from that population. The "confidence level" controls the precision of your inference: it corresponds to the probability that the interval includes (or "covers") the population parameter. The "error percentage" is 100 minus the confidence level; it corresponds to the chance that the "true" parameter value is not within the confidence interval.

Choosing a higher confidence level yields less chance of error, but also a less precise (i.e., wider) interval. Two conventional choices for confidence levels are 95 and 99; each yields high confidence that the interval does include the true parameter.

To get confidence intervals for a mean via the "Stata Command" window, issue the following command:

```
ci means <varname>,level(##)
```

where you fill in the variable name of interest to you in place of "varname" and designate your selected confidence level in place of "##". (If you do not use the "level" option, Stata will assume the conventional 95% confidence level.)

For example, the following command produces 95% confidence levels for the mean of variable "tvhours" (number of hours per day watching TV) in data set GSS2016:

```
ci means tvhours
```

After issuing this command, the following output appears in the "Stata Results" window:

```
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     tvhours |      1,883    3.031333    .0647707        2.904303    3.158363
```

We are 95% confident that the interval (2.90, 3.16) includes the mean of the number of hours of television watched per day within the population of English- or Spanish-speaking US adults in 2016. There is a 5% chance that this inference is incorrect, i.e. that this range of values does not include the population mean.

If we want greater certainty than 95% (the default), we can raise the confidence level by adding the "level" option and requesting higher confidence, in this instance 99%:

```
ci means tvhours,level(99)
```

which yields the following report:

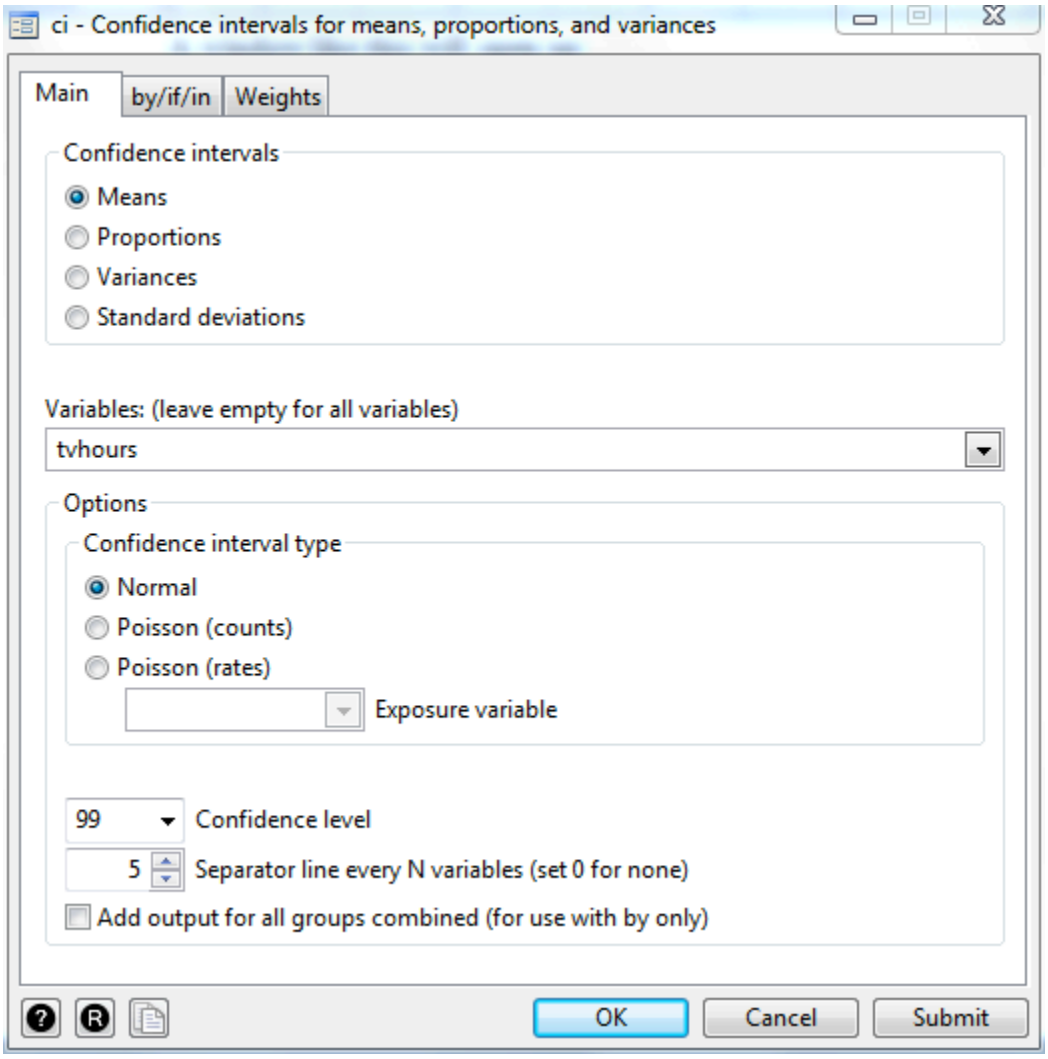
```
    Variable |        Obs        Mean    Std. Err.       [99% Conf. Interval]
-------------+---------------------------------------------------------------
     tvhours |      1,883    3.031333    .0647707        2.864325    3.198341
```

We are 99% confident that the interval (2.86, 3.19) contains the mean number of hours of television viewing per day within the population of English- or Spanish-speaking US adults in 2016. There is only a 1% chance that this inference is incorrect. This interval seems only a little wider than the 95% interval, but its estimate is somewhat less exact. The lowered precision of the estimate compensates for the reduced chance of error.

If you prefer to use the Stata menus to do this, proceed as follows:

click on "Statistics"
click on "Summaries, tables, & tests"
click on "Summary and Descriptive Statistics"
click on "Confidence Intervals"

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



Fill in the variable name of interest to you in the "Variables:" box. Leave the button for "Normal" checked under "Options". Select your "Confidence level" using a drop-down list in the box at the lower left of the screen; the level will default to 95 if you do not alter it there.

Then hit "OK" and a report like the one shown earlier (for the 99% interval) will appear.

If you want confidence intervals for the proportion of a population that has some outcome, you must work with a dichotomous variable that is coded 1 if someone has the outcome of interest, and 0 otherwise. You may have to recode your variables in order to do this (see separate handout on how to recode). For example, to code a categorical variable that measures whether or not LGBTQ individuals have the right to marry (variable "marhomo," 1=strongly agree…5=strongly disagree) as a proportion, the recoding and variable labeling commands might be

```
recode marhomo (1/2=1 "Favor")(3/5=0 "Neutral or
oppose"),gen(marhomo_r)

label variable marhomo_r "Favorable view toward gay marriage"
```

(Note: enter the commands in Stata on one line. "oppose…" is displayed on the second line above due to Microsoft Word formatting.).

Then you issue the following command:

```
ci proportions <varname>,level(##) wald
```

Again, if you omit the "level" option, Stata will construct a 95% confidence interval.

For the gay marriage example a 95% confidence interval for the proportion favorable is obtained as follows:

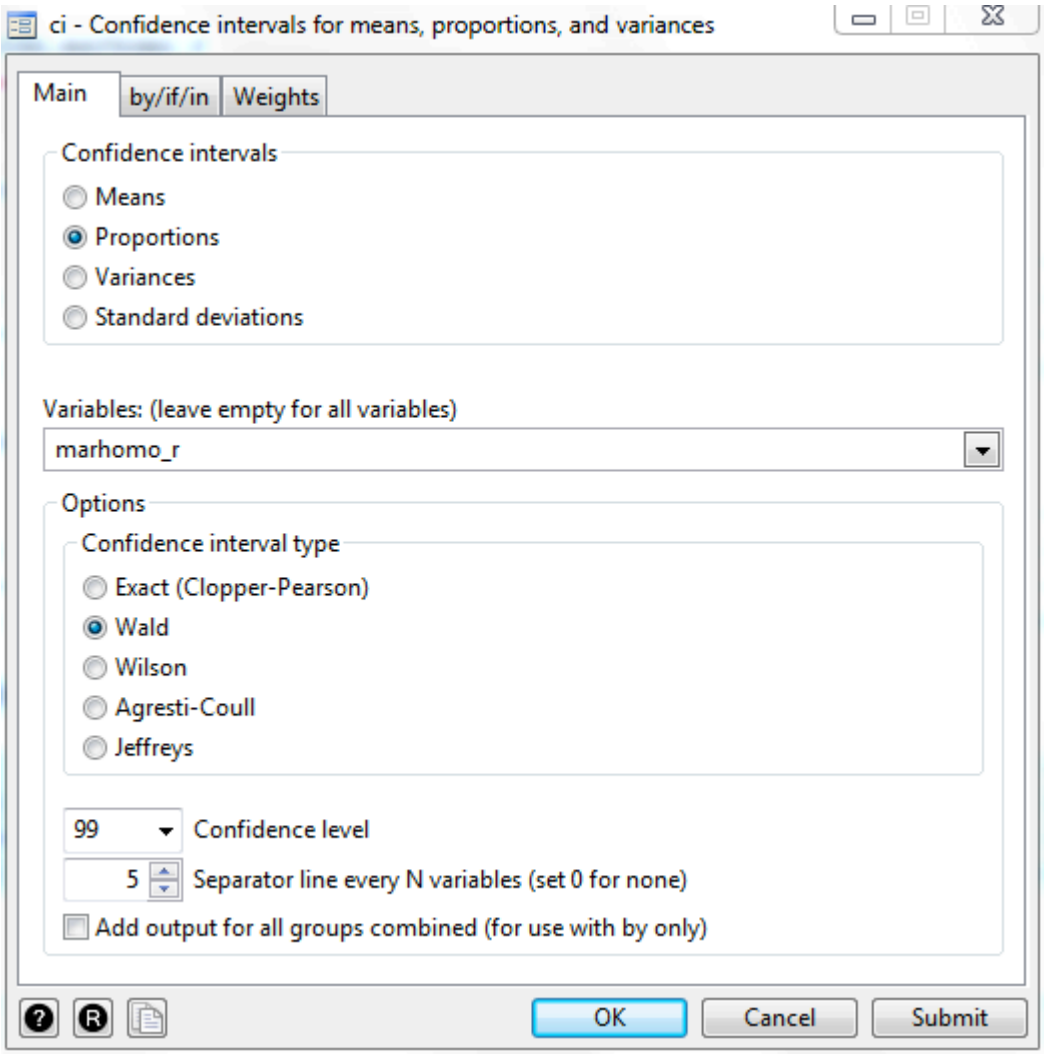
```
. ci proportions marhomo_r, wald

                                                        -- Binomial Wald ---
    Variable |        Obs  Proportion    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
   marhomo_r |      1,862    .5929108    .0113854        .5705958    .6152259
```

We are 95% confident that the interval (0.571, 0.615) includes the proportion favorable to gay marriage within the population of English- or Spanish-speaking US adults in 2016.

To accomplish this via the Stata menus, proceed as follows:

click on "Statistics"
click on "Summaries, tables, & tests"
click on "Summary and Descriptive Statistics"
click on "Confidence Intervals"



In the window that opens, first check "Proportions"; doing this will change the content of the rest of the window. Enter the dichotomous (0/1) variable of interest in the "Variables:" box, and click the "Wald" button. If you want a level of confidence other than 95%, select it in the "Confidence level" box (a 99% interval was requested here). Then click "OK" to obtain this report:

```
                                                        -- Binomial Wald ---
    Variable |        Obs  Proportion    Std. Err.       [99% Conf. Interval]
-------------+------------------------------------------------------------------
   marhomo_r |      1,862    .5929108    .0113854        .5635839    .6222378
```

This indicates that there is a 99% chance that the interval from 0.564 to 0.622 contains the proportion favoring gay marriage in the population of all U.S. adults.