Custom Stata Commands for Semi-automatic Confidentiality Screening of Statistics Canada Data

Jesse McCrosky, M.Math.

Thursday, July 26th, 2012
2012 Stata Conference, San Diego

Outline

- Overview
- The Research Data Centres (RDCs)
- Why Stata?
- Why Custom?
- The Commands
- Conclusions

Overview

- Content is specific to Statistics Canada Research Data Centres, but...
- Demonstrate how Stata can be customized for specific applications and environments
- Present commands that may be useful in a variety of situations and may inspire new ideas

The Research Data Centres

- Provide secure access to confidential microdata
- Cooperation between academic researchers and Statistics Canada
- Balance:
  - Promoting and facilitating research
  - Protecting confidentiality of respondents

RDC Release Process

- Researchers must work inside RDC with no network access or removable media

- Only aggregate results can be released, i.e. model output, frequency tables, and other descriptive statistics

- Before release, results must be vetted by an analyst to ensure no confidential results are released

Confidential Results

- Policies vary between different surveys, but, for example:
  - Results must be weighted and based on at least 5 unweighted respondents
  - Certain types of output must be rounded
  - Dominance and homogeneity of dollar value variables

Why Stata?

- RDCs provide SPSS, SAS, and Stata, as well as other statistical software packages
- Internally, Statistics Canada uses mostly SAS, although Stata is gaining traction
- Most surveys provide bootstrap weights for robust variance estimation
  - Stata "svy" prefix is extremely useful
- Ease of researcher support

Why Custom?

- Verifying that results meet confidentiality requirements generates work for both researchers (RDC users) and analysts (RDC employees)
- Vetting process can be error prone, especially with very large amounts of output
- Save time and decrease likelihood of errors

The Commands

- Frequency tables
- Model output
- Pseudo min/max
- Dominance and homogeneity

Frequency Tables

- Three enhancements:
  - Enforce correct use of weight variable
  - Reporting minimum unweighted cell size
  - Automatic rounding

Enforce Correct Use of Weight Variable

- Supplied master weights are probability weights but can be interpreted as pseudo-frequency weights for population frequency
- Can be a little tricky:
  - `tab [pw=wtvar]` doesn't work
  - `tab [iw=wtvar]` works, but is not ideal conceptually
  - `svyset [pw=wtvar]` then `svy: tab` works for proportions only
  - `table [pw=wtvar]` works for frequencies only
- Custom command can hide this complexity

Reporting Minimum Unweighted Cell Size

- Want to produce weighted table but need to know minimum unweighted cell size to determine releasability

- Issue of zero-cells

- Normally researchers produce both weighted and unweighted versions of tables

  - Requires extra time and results in more output produced

Example Weighted and Unweighted Tables

Unweighted		Gender	
		Male	Female
IBS	Yes	645	893
	No	221	196

Weighted		Gender	
		Male	Female
IBS	Yes	3,749	5,982
	No	40,398	42,587

Releasability only depends on unweighted table!

Automatic Rounding

- Under some situations, frequencies must be rounded
- Different rounding algorithms may be allowed in different situations

Table Example

```
. stctab gender agegrp

           |
           |          Gender
 Age Group |      Male     Female |     Total
-----------+----------------------+----------
     15-24 |        34         39 |        73
     25-34 |        44         39 |        83
     35-44 |        32         27 |        59
     45-54 |        33         28 |        61
       55+ |        19         22 |        41
-----------+----------------------+----------
     Total |       162        155 |       317
```

```
. stctab gender agegrp, weighted(wts_m) vet(5)

           |
           |          Gender
 Age Group |      Male     Female |     Total
-----------+----------------------+----------
     15-24 |     3,329      4,000 |     7,329
     25-34 |     4,827      3,872 |     8,699
     35-44 |     2,901      2,890 |     5,791
     45-54 |     3,698      2,607 |     6,305
       55+ |     2,010      2,098 |     4,108
-----------+----------------------+----------
     Total |    16,765     15,467 |    32,232

Table satisfies minimum unweighted cell size
of 5.
```

Table Example

```
. stctab empstat deceased

           |
           |   Employed?
 Deceased? |      Yes       No |      Total
-----------+-------------------+-----------
       Yes |        0       44 |         44
        No |      137        4 |        141
-----------+-------------------+-----------
     Total |      137       48 |        185
```

```
. stctab empstat deceased, weighted(wts_m) simpleround(10)

           |
           |   Employed?
 Deceased? |      Yes       No |      Total
-----------+-------------------+-----------
       Yes |        0      800 |        800
        No |    4,360      200 |      4,560
-----------+-------------------+-----------
     Total |    4,360    1,000 |      5,360
```

```
. stctab empstat deceased, weighted(wts_m) vet(5)

           |
           |   Employed?
 Deceased? |      Yes         No |      Total
-----------+---------------------+-----------
       Yes |        0        802 |        802
        No |    4,356        203 |      4,559
-----------+---------------------+-----------
     Total |    4,356      1,005 |      5,361
```

Minimum unweighted non-zero cell size is too small for release.

Table Example

```
. stctab stress, vet(15)

Self-report |
     Stress |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |           2        2.90        2.90
          2 |           8       11.59       14.49
          3 |          30       43.48       57.97
          4 |          18       26.09       84.06
          5 |          11       15.94      100.00
------------+-----------------------------------
      Total |          69      100.00

Minimum unweighted non-zero cell size is too small for release.
```

Model Output

- One enhancement:
  - Indicate releasability
- Possible additional enhancements:
  - Identify “risky” models: saturated or bivariate dichotomous

Indicate Releasability

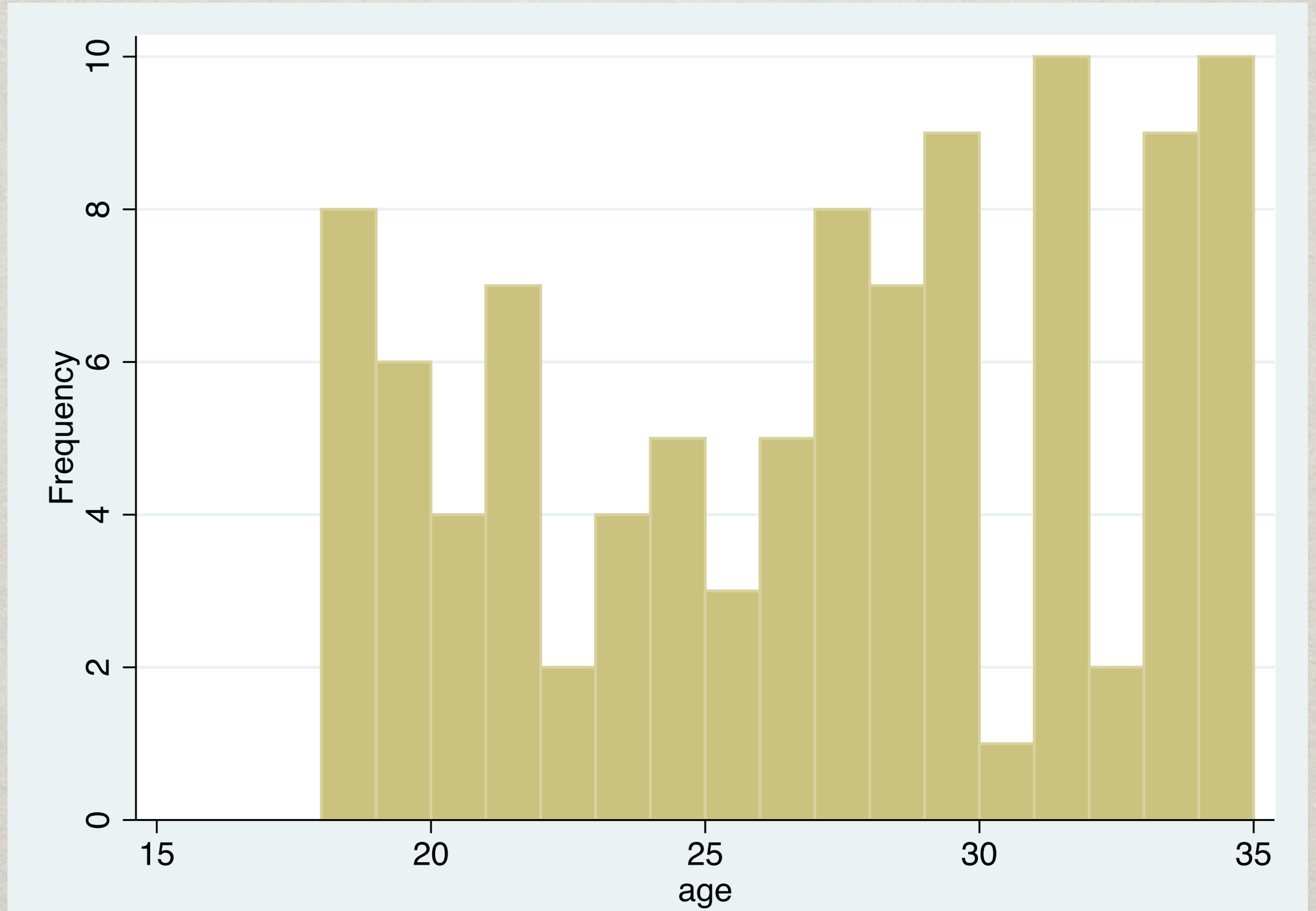
```
. stclogit incidence sex stress age depression dc, vet(5)
                                                Number of obs      = sufficient
                                                Population size    =  106658541

------------------------------------------------------------------------------
             |   Observed
   incidence |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -2.812405   .6252287    -4.50   0.000    -4.037831    -1.58698
      stress |   .4471138   .2760224     1.62   0.105    -.0938801    .9881077
         age |  -.5859508   .2840305    -2.06   0.039     -1.14264   -.0292611
  depression |  -9.551872   2.978553    -3.21   0.001    -15.38973   -3.714016
          dc |   .0665457   .0822958     0.81   0.419     -.094751    .2278425
       _cons |    5.55052   5.347551     1.04   0.299    -4.930488    16.03153
------------------------------------------------------------------------------
```

Don't display actual number of observations as many surveys disallow weighted and unweighted version of the same frequency

Pseudo Min/Max

- Problem: minimum and maximum are unreleasable by definition as they're based on a single respondent
- Solution: use extreme percentiles instead, i.e. 95th and 5th
- For minimum cell-size of *m* and sample size *n*, may release (*m*/*n* * 100)th and (100 - *m*/*n* * 100)th percentiles or better
  - If minimum and maximum values are not unique, may be more liberal

Min/Max may Releasable



Pseudo Min/Max Example

```
. stcminmax visits, vet(5)


 Visits to GP |
--------------+-----------
         obs  |       456
         min  |         0
   99th perc. |        14
-------------------------


. stcminmax visits, p(1)


 Visits to GP |
--------------+-----------
         obs  |       456
    1st perc. |         0
   99th perc. |        14
-------------------------
```

Dominance and Homogeneity

- In some cases, results involving dollar value variables are subject to dominance and homogeneity requirements:

- Dominance: maximum income must not be more than x% of the total income of all respondents in the sub-population

- Homogeneity: range of incomes must be at least x% of the maximum income in the sub-population

Dominance and Homogeneity Example

```
. stcdhtable income ethnic gender


Variable income meets dominance and homogeneity requirements over
all values of “ethnic” and “gender”.


. stcdhmodel income ethnic gender


Variable income meets dominance and homogeneity requirements for
sub-population non-missing for specified variables.
```

- Could be integrated into other commands

Conclusions

- Writing custom Stata commands provides a fairly simple way to create tools to meet specific output requirements

- Can save significant time for researchers and analysts

- Still, commands are an aid only, and analyst must still carefully examine output and collaborate with researcher in ensuring that confidentiality is respected

Thank you

- Thank you for your time!
- Feel free to follow up by e-mail
  - Jesse McCrosky
  - mccrosky@gmail.com
- Code for commands described and an article (to be published in the Statistics Canada Internal Technical Bulletin) will be available soon
- Thanks to the University of Saskatchewan Faculty of Graduate Studies and Research for their support of my participation in this conference
- Questions?