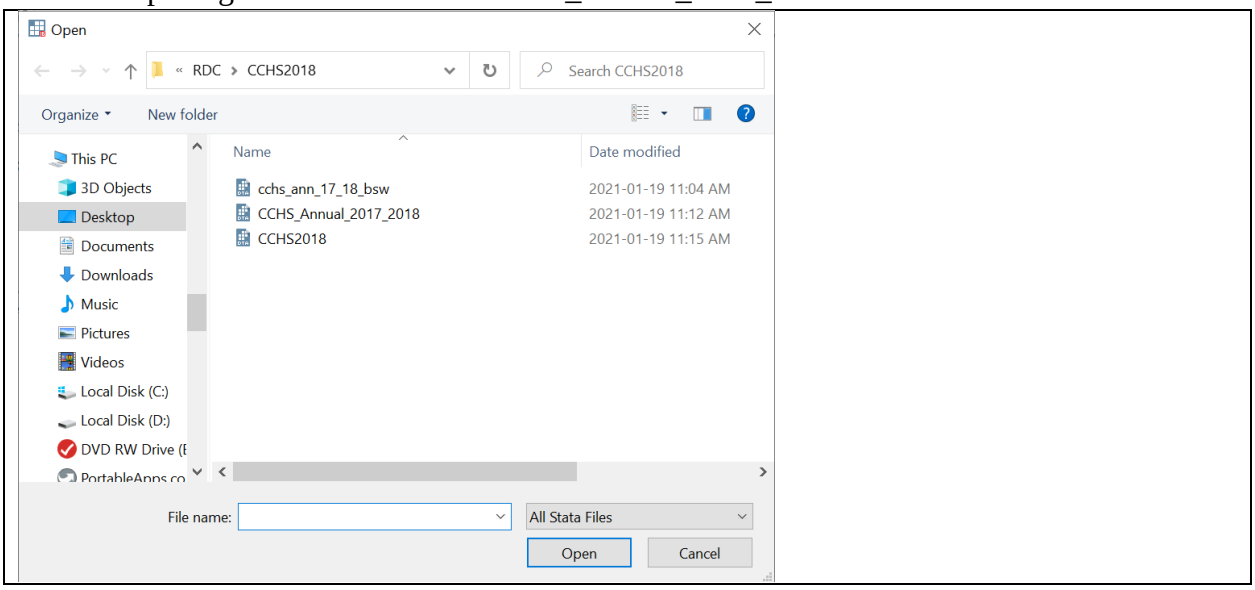
Analyzing Complex Sample Surveys with STATA™

Data-File: The data for this example is taken from the Canadian Community Health Survey administered by Statistics Canada in 2017-18 (CCHS, 2018). There is a Public User Microdata File (PUMF) version of this data-set, readily available at most Canadian universities under the Data Liberation Initiative (DLI). The data file is provided in a Stata data file format (CCHS_Annual_2017_2018.sta), and also has available a companion bootstrapping weights file (cchs_ann_17_18_bsw.sta). Always check the **User Guide** that comes with a given Statistics Canada data-set for specific details on how to analyze it.

Table 1: Opening the Stata data file: CCHS_Annual_2017_2018.sta



Person-level Weight Variable: In order to correct for any nonresponses and disproportionate sampling specific to a given sample, there is a variable in the data file for individual person-level or cohort weights, called **WTS_M**. This variable is scaled to the population size, and basically represents the number of people who are in the same population strata for a given case (i.e., having the same values for key demographic variables such as geographical location, age, gender, etc.). In order to analyze first-order statistics (i.e., proportions & means), this weight variable should be applied in order to make the results more representative of the population.

Using the Person-Level Weight for Nonresponses: For frequency tables, weighting of cases to correct for nonresponses can be set on using the importance weight (**iweight**) subcommand. Analyze the variable **DHH_SEX** (for sex/gender) using the Stata **tabulate** command, with weighting turned off and on [iweight = **WTS_M**], to determine the effect of controlling for the sampling bias due to nonresponses (**ON**).

Richard MacLennan, PhD
Professor of Psychology, U. of Regina
Academic Director, Regina RDC

Table 2: Frequency table analysis of **DHH_SEX** with weighting of cases turned OFF and ON.

With Weighting of Cases turned **OFF**

```
. tabulate DHH_SEX

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |     52,402       46.25       46.25
     Female |     60,888       53.75      100.00
------------+-----------------------------------
      Total |    113,290      100.00
```

With Weighting of Cases turned **ON**

```
. tabulate DHH_SEX [iweight = WTS_M]

        Sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male | 15432670.6       49.35       49.35
     Female | 15841701.4       50.65      100.00
------------+-----------------------------------
      Total | 31,274,372      100.00
```

As can be seen in the first analysis with weighting turned **OFF,** these results are reported relative to the sample size (113,290), instead of relative to the population size (31,274,372 rounded) as reported in the second analysis with weighting turned **ON**. In addition, in the second analysis the percentages have changed to control for ***nonresponses*** in some strata (relative to what would be expected in the population). For instance, in many surveys males are under-represented in samples because they tend to volunteer to participate less often (only 46.25% in this sample). However, with weighting turned **ON**, this has been corrected to the relative proportion of males in the population (approximately 49.35% based upon both expectations from genetic theory and also from projections from the 2016 general population census), so that as a group they are not unduly affected by the lower response-rate by some members of their own strata.

Disproportionate Sampling: Another use of person-level weights is to correct for sampling bias due to ***disproportionate sampling***. In some cases, Statistics Canada may *intentionally* oversample some groups, especially in the case of small groups (say for geographic locations with smaller populations, elderly persons, people with disabilities, etc.). This is order to not entirely miss some of these smaller groups, such as might occur if random sampling was employed and it merely was left to chance. On the other hand, if smaller groups are over-sampled on purpose then larger groups hence will be under-sampled by implication. In order to correct for this disproportionate sampling, person-level weights can be applied afterwards during the analyses to adjust the relevant group proportions in a sample so that they look more like the actual group proportions in the population.

The following pie-charts and frequency tables show the relative proportion of respondents in the CCHS (2018) broken down by provinces and territories in the sample data (***unweighted***), and the same estimated proportions in the population data (***weighted***). In comparing the two sets of results, you can see that almost all of the provinces and territories are over-sampled in the sample data compared to the population data, with the exception of Ontario and Quebec (which are under-sampled). Hence there is a need to ***right-size*** these proportions by using weighting of cases to make the results more comparable to the population. If you conceptualize the weighted cases pie-chart as a dart board, you can see how it would be virtually impossible to "hit" some of the smaller groups (say PEI or the northern territories) if just random sampling was employed.

Table 3: Pie-charts & Frequency Tables for Number of Respondents per Province or Territory

Sample Data:

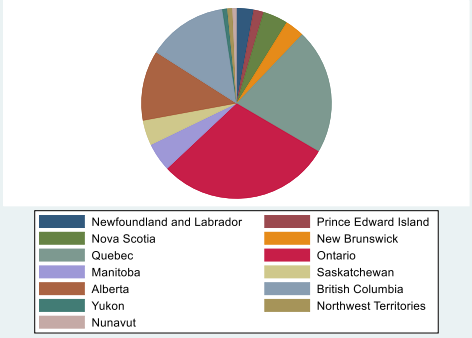
Unweighted Frequency Table for Province

```
. tabulate GEO_PRV

  Province of |
   respondent |      Freq.     Percent        Cum.
--------------+-----------------------------------
           NL |      3,291        2.90        2.90
           PE |      1,928        1.70        4.61
           NS |      4,811        4.25        8.85
           NB |      3,706        3.27       12.12
           QC |     24,125       21.29       33.42
           ON |     33,511       29.58       63.00
           MB |      5,481        4.84       67.84
           SK |      4,835        4.27       72.11
           AL |     13,486       11.90       84.01
           BC |     15,410       13.60       97.61
           YK |        956        0.84       98.46
           NT |        980        0.87       99.32
           NU |        770        0.68      100.00
--------------+-----------------------------------
        Total |    113,290      100.00
```

Unweighted Pie Chart for Province

```
. graph pie, over(GEO_PRV)
```



Population Data:

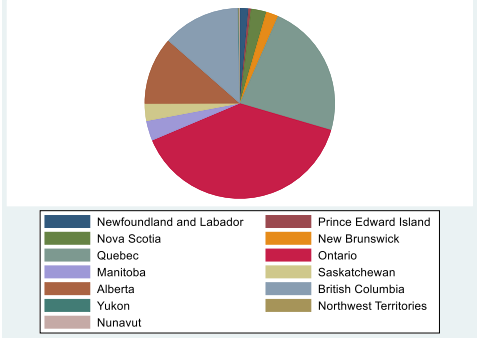
Weighted Frequency Table for Province

```
. tabulate GEO_PRV [iweight = WTS_M]

  Province of |
   respondent |      Freq.     Percent        Cum.
--------------+-----------------------------------
           NL | 459,931.23        1.47        1.47
           PE | 129,507.12        0.41        1.88
           NS | 821,775.63        2.63        4.51
           NB | 648,259.14        2.07        6.59
           QC |  7,173,161       22.94       29.52
           ON | 12235212.5       39.12       68.64
           MB |  1,060,834        3.39       72.04
           SK | 917,473.86        2.93       74.97
           AL |  3,590,200       11.48       86.45
           BC |  4,141,320       13.24       99.69
           YK |  32,734.48        0.10       99.80
           NT |  35,445.51        0.11       99.91
           NU |  28,517.86        0.09      100.00
--------------+-----------------------------------
        Total | 31,274,372      100.00
```

Weighted Pie Chart for Province

```
. graph pie [pweight = WTS_M], over(GEO_PRV)
```

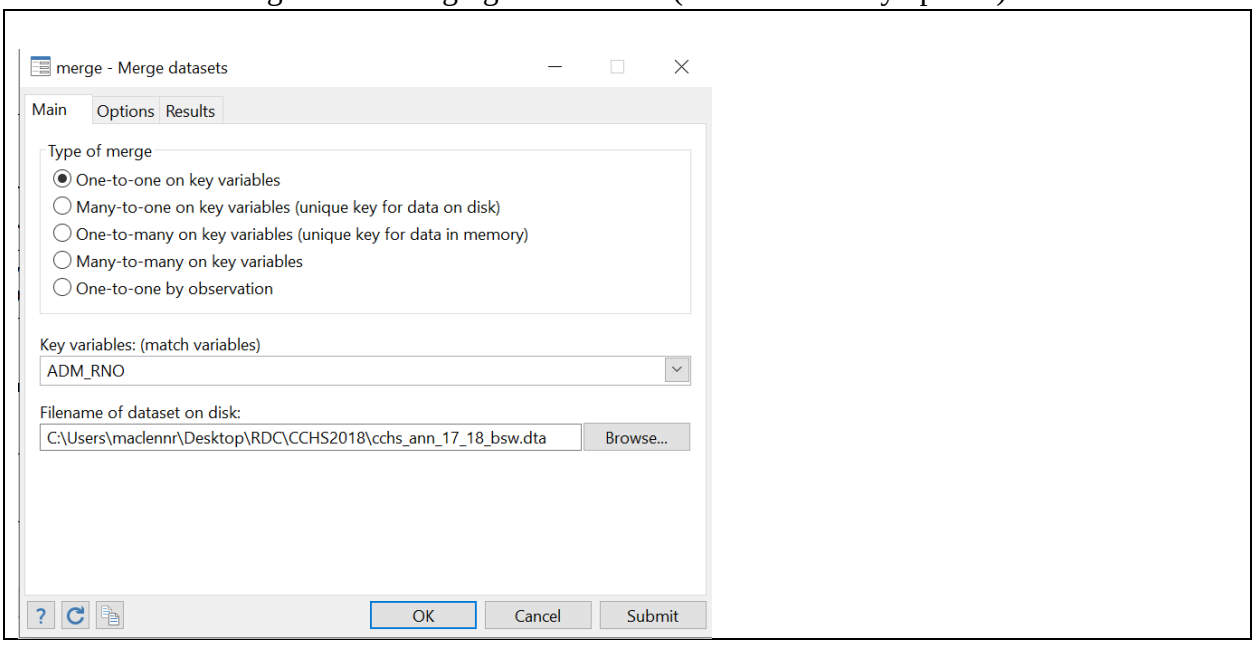


Bootstrapping Weights: Although individual person-level weights can help improve the analyses for first-order statistics (i.e., means & proportions) in complex sample surveys, they are ineffective for higher-order statistics (such as variances, standard errors, correlations, etc.). In this case, to take into account the complex sampling procedures in a given survey (i.e., multistage, stratification, & clustering), bootstrapping weights must be applied for these advanced statistical analyses. Bootstrapping is an intensive re-sampling procedure (with replacement of cases after the selection of each case), in order to empirically simulate the distribution of statistical results for the population from the given sample. For Statistics Canada data-sets such as the CCHS 2017-18, they provide a file (cchs_17_18_bsw_.sta) with an additional 1,000 variables that employ bootstrap weights to adjust for the complex sampling design in the given survey (e.g., multistage, stratification, & clustering).

Merging the Stata data file (CCHS_Annual_2017_2018.sta) and the bootstrapping weights file (cchs_17_18_bsw.sta): In order to apply the bootstrapping weights, first the two files for data and bootstrapping weights must be merged together (CCHS_Annual_2017_2018.sta & cchs_17_18_bsw.sta, respectively). To be merged, the two files must have a common indexing variable, so that their cases can be matched. For the CCHS 2017-18 survey, both files have an indexing variable in common, the administration record number (**ADM_RNO**).

With the CCHS_Annual_2017_2018.sta file already opened in Stata, you can merge the bootstrapping weights file cchs_17_18_bsw.sta by pulling down the Data menu, clicking on Combining datasets, and selecting the Merge two datasets option to open the following dialog:

Table 4: Stata Dialog box for merging two datasets (with one already opened).



Richard MacLennan, PhD
Professor of Psychology, U. of Regina
Academic Director, Regina RDC

Use the Merge Datasets dialog box to identify the key matching variable (**ADM_RNO**) and the new file to be merged (cchs_17_18_bsw.sta) with the currently opened Stata data file (CCHS_Annual_2017_2018.sta.dta).

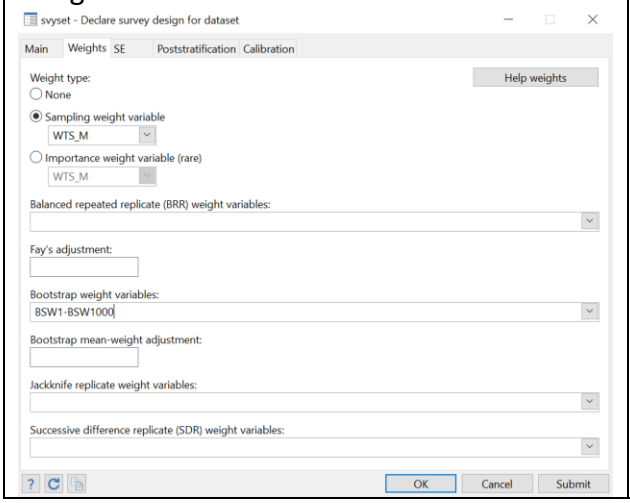
When the two files are merged, you will now see 1,000 new variables for bootstrapping weights at the end of the variable list (BSW1 to BSW1000). [BSW stands for "BootStrapping Weights".] Once you confirm that the two files have been successfully merged, it would be a good idea now to save the merged file under a new name (something like: CCHS2018.dta).

Bootstrap Mean-Weight Adjustment: Sometimes due to the bootstrapping method employed in a given data-set, it is possible to obtain negative bootstrap weights (an anomalous result). In this case, bootstrap weights need to be transformed to always be positive. Hence to correctly estimate sampling errors necessary for statistical hypothesis testing and confidence intervals, estimates have to be modified by a constant called the **Bootstrap Mean-Weight Adjustment** (the Stata software package will automatically use this value to undertake the calculation internally to correctly estimate sampling errors, without requiring the user to do so separately). *Fortunately, the bootstrap mean-weight adjustment is **unnecessary** for the CCHS2018 dataset.* [**Please Note:** This mean-weight adjustment may be different for different data-sets & software].

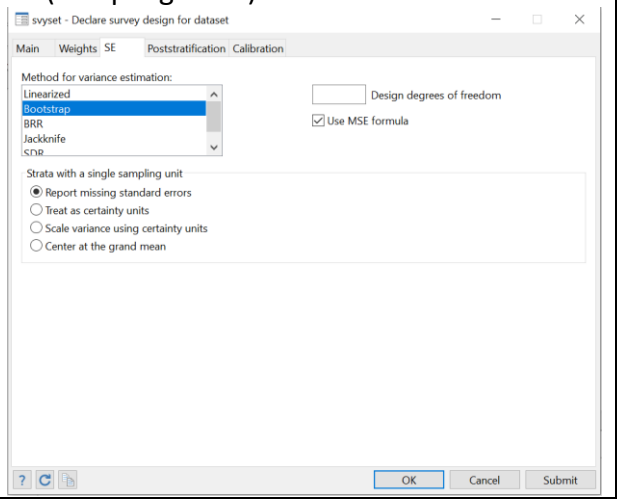
To set-up the parameters for bootstrap estimation of sampling error, pull-down the Statistics Menu, select the Survey data analysis item, select the Set-up & utilities item, then Declare Survey design for dataset option. Fill-out the Weights and SE Tab dialogs as follows:

Table 5: Setting Parameters for Bootstrap Estimation in Stata (no Bootstrap mean-weight adjustment necessary here for the CCHS, 2018).

Weights Tab



SE (Sampling Error) Tab



Richard MacLennan, PhD
Professor of Psychology, U. of Regina
Academic Director, Regina RDC

Equivalent Stata syntax for SVYSET command:

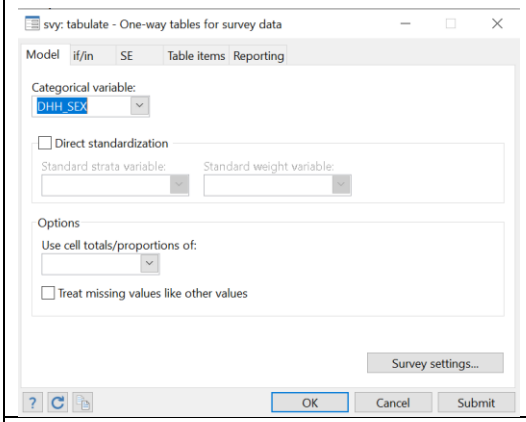
```
. svyset _n [pweight=WTS_M], bsrweight(BSW1-BSW1000) vce(bootstrap) mse singleunit(missing)

      pweight: WTS_M
          VCE: bootstrap
          MSE: on
    bsrweight: BSW1 .. BSW1000
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

A. Creating a frequency table in Stata taking into account complex sampling:
Once the survey design has been declared, to set-up the parameters for frequency tabulation of the **DDH_SEX** variable taking into account the complex sample design, pull-down the Statistics Menu, select the Survey data analysis item, select the Tables item, then select the One-way table option. Fill-out the Model tab, SE tab, Table Items tab, & Reporting tab dialogs as follows:

Table 6: Set-up for tabulation of the **DDH_SEX** variable, taking into account complex sampling

Model tab



SE (sampling error) tab

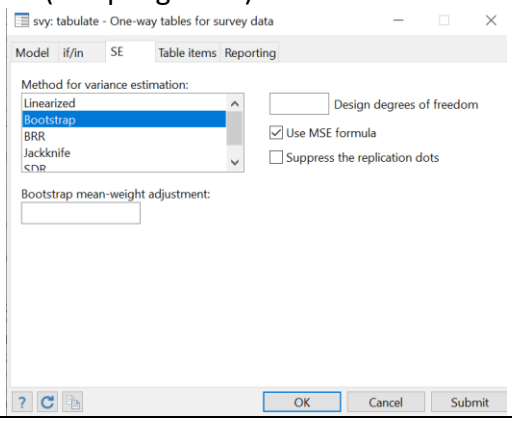
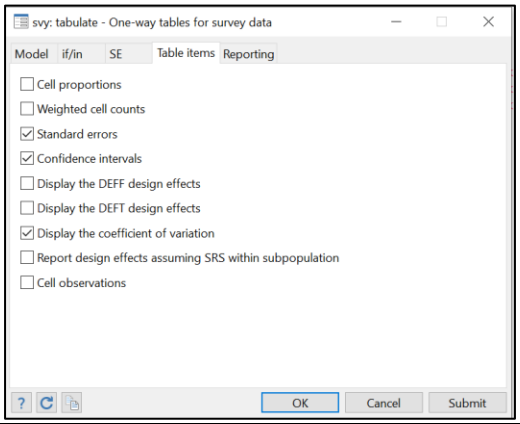
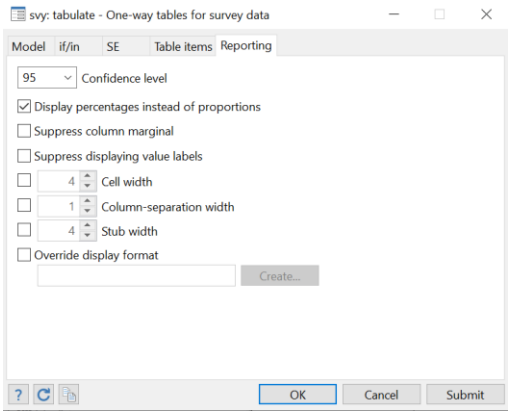


Table Items tab



Reporting tab



Richard MacLennan, PhD
Professor of Psychology, U. of Regina
Academic Director, Regina RDC

Table 7: Results for tabulation of the **DDH_SEX** variable taking into account the complex sample

```
. svy bootstrap, mse : tabulate DHH_SEX, se ci cv percent
(running tabulate on estimation sample)

                                            Number of obs     =     113,290
                                            Population size   =  31,274,372
                                            Replications      =       1,000

------------------------------------------------------------------------------
         Sex | percentage          se          cv          lb          ub
-------------+----------------------------------------------------------------
        Male |      49.35       .0069       .0139       49.33       49.36
      Female |      50.65       .0069       .0136       50.64       50.67
             |
       Total |         100
------------------------------------------------------------------------------
  Key:  percentage  =  cell percentage
        se          =  bootstrap standard error of cell percentage
        cv          =  coefficients of variation of cell percentage
        lb          =  lower 95% confidence bound for cell percentage
        ub          =  upper 95% confidence bound for cell percentage
```

In the frequency table analysis output, the reader may notice that percentages for Males and Females are exactly the same as in the previous analysis with weighting of cases turned **ON**. In addition, this table now reports for each percentage: the standard error (SE), the coefficient of variation (CV), and the lower & upper bound of the 95% confidence interval (LB & UB, respectively). These statistics are ***only*** available when bootstrapping weights are applied.

The SE is used to derive both the 95% confidence interval (CI = percentage ± 1.96 × SE), and the coefficient of variation (CV = SE/percentage). As it turns out, none of the 95% CI's overlap with each other, indicating that the percentages for Males and Females are significantly different from each other. In addition, all of the CV's are less than .166 indicating that the percentages are acceptable to report according to Statistics Canada's standards (i.e., CV < .166 is acceptable, .166 ≤ CV ≤ .333 is marginal, and CV > .333 is unacceptable to report the percentage).

B. Performing a t-test for means in Stata taking into account complex sampling:

Step 1. Setting-up the Survey Design by using the *svyset* command in Stata

```
. svyset _n [pweight=WTS_M], bsrweight(BSW1-BSW1000) vce(bootstrap) mse
singleunit(missing)

      pweight: WTS_M
          VCE: bootstrap
          MSE: on
    bsrweight: BSW1 .. BSW1000
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

Step 2. Use the *svy : mean* command in Stata to find the mean heights of men and women (in metres), and the associated coefficient legends (needed for the next command)

```
. svy : mean hwtdghtm, over(DHH_SEX) coeflegend
(running mean on estimation sample)

Bootstrap replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................  1000
```

[**Note:** The above graphic shows the progress of the bootstrapping for 1000 replications.]

```
Survey: Mean estimation          Number of obs    =     107,959
                                 Population size =  29,644,622
                                 Replications     =       1,000

------------------------------------------------------------------------------
                     |   Observed
                     |       Mean  Legend
---------------------+--------------------------------------------------------
c.hwtdghtm@DHH_SEX   |
              Male   |     1.7655  _b[c.hwtdghtm@1bn.DHH_SEX]
            Female   |   1.628015  _b[c.hwtdghtm@2.DHH_SEX]
------------------------------------------------------------------------------
```

It should be noted that Stata retains information in memory from some analyses, which can then be used in further statistical analyses (referred to as *post-estimation).* The coefficient *legend* parameter is the way that Stata keeps track of where this information is stored, as will be seen in the following analysis. [The *legend* parameters for males and females can be cut-&-pasted into the next command.] A post-estimation analysis will only work if it immediately follows the required preliminary analysis, which stores the relevant information.

Step 3. Using the *lincom* (linear combination) command in Stata to test if the mean height for men *minus* the mean height for women is significantly different from 0.0 (using *legend for largest mean – legend for smallest mean***), with df = N – 2 (df = 107957 in this case)**

```
. lincom _b[c.hwtdghtm@1bn.DHH_SEX] - _b[c.hwtdghtm@2.DHH_SEX], df(107957)

 ( 1)  c.hwtdghtm@1bn.DHH_SEX - c.hwtdghtm@2.DHH_SEX = 0

------------------------------------------------------------------------------
        Mean |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1374859   .0007939   173.19   0.000       .13593    .1390419
------------------------------------------------------------------------------
```

Note that the coefficient (.1374859) is equal to the difference in mean heights between men and women (in metres), and the 95% confidence interval around this estimate is .13593 (lower bound) to .1390419 (upper bound), based upon bootstrap estimation of the variance. The 95% CI is approximately equal to the coefficient (.1374859) ± 1.96 × its standard error (.0007939). It should be noted that the standard error is actually estimated using bootstrapping over 1000 replications. The value of the t-test (173.19) is also equal to the coefficient (.1374859) divided by its respective standard error (.0007939). The value of the t-test is positive in this case because the order of subtraction was *largest mean – smallest mean*. The probability of getting this value of the t-test or larger by chance if the null hypothesis was true (i.e., that there is no difference in mean heights) is only $p < .001$, so we would declare that the mean height of men is significantly larger than the mean height of women for the Canadian population (by a factor of .1374859 metres or 13.74859 cms.)

C. Finding a correlation in Stata taking into account complex sampling:

Step 1. Setting-up the Survey Design using the *svyset* command in Stata

```
. svyset _n [pweight=WTS_M], bsrweight(BSW1-BSW1000) vce(bootstrap) mse
singleunit(missing)

      pweight: WTS_M
          VCE: bootstrap
          MSE: on
    bsrweight: BSW1 .. BSW1000
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
```

Step 2. Use the *svy : regress* command in Stata to find the R-squared value between height and sex

```
. svy : regress hwtdghtm DHH_SEX
(running regress on estimation sample)

Bootstrap replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................  1000
```

[**Note:** The above graphic shows the progress of the bootstrapping for 1000 replications.]

```
Survey: Linear regression                       Number of obs     =     107,959
                                                Population size   =  29,644,622
                                                Replications      =       1,000
                                                Wald chi2(1)      =    29994.03
                                                Prob > chi2       =      0.0000
                                                R-squared         =      0.4585

------------------------------------------------------------------------------
             |   Observed   Bstrap *
    hwtdghtm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     DHH_SEX |  -.1374859   .0007939  -173.19   0.000    -.1390419     -.13593
       _cons |   1.902986   .0013424  1417.62   0.000     1.900355    1.905617
------------------------------------------------------------------------------
```

The value of R-squared in this case is .4585. Therefore, sex accounts for approximately 45.85% of the variance in height. Note that the absolute value of the Z-test (173.19) is the exact same as the value of the t-test in the previous analysis for the difference in mean heights of men and woman. This is because for very large samples such as in this case (N = 107,959), the Gaussian normal (Z) distribution and Student's t-distribution are virtually identical. Indeed, if you omit the *df(107957)* subcommand in the previous *lincom* analysis, Stata will do a Z-test instead of a t-test. The absolute value of the regression coefficient for sex and its associated standard error (based upon bootstrapping estimation as highlighted in blue) are also the same as in the previous analysis. In addition, the value of the overall Wald χ^2 test statistic (with df = 1 for one independent variable) happens to equal the absolute value of the Z-statistic for sex squared, within rounding. [The Wald χ^2 test is also equal to the value of the F-test of an ANOVA summary table for the regression analysis, if you set the design *dof* = 107957].

Step 3. Using the *display* command in Stata to find the value of Z-squared (or Wald's χ^2)

```
. display 173.19*173.19
29994.776
```

Step 4. Using the *display* command in Stata to find the square-root of the R-squared value or the correlation between height and sex

```
. display sqrt(.4585)
.67712628
```

Therefore the simple correlation between height and sex is -.67712628, which is statistically significant with Wald's χ^2 = 29994.03, p <.0001 from the regression analysis. Since the regression coefficient is negative, the sign of the correlation coefficient would also be negative. However, the sign is arbitrary in this case because it is dependent upon the coding of the dichotomous variable (i.e., male = 1, female = 2).

Table 8: Menu of Other Analyses for Complex Sample Surveys in Stata (highlighted in yellow):