Complex Samples

Richard N. MacLennan
Professor, Psychology
University of Regina &
Academic Director, Statistics Canada
Regina Research Data Centre (RDC)

Impetus (Disciplines of RDC Researchers)

DISCIPLINE	RESEARCHERS
ECONOMICS	641
PUBLIC HEALTH / EPIDEMIOLOGY	362
SOCIOLOGY / ANTHROPOLOGY	286
PSYCHOLOGY / PSYCHIATRY	108
GEOGRAPHY / URBANISM	105
MATHEMATICS / STATISTICS	78
MEDICINE	66
INDUSTRIAL RELATIONS	61
DEMOGRAPHY	60

DISCIPLINE	RESEARCHERS
NUTRITION	47
POLITICAL SCIENCE	28
EDUCATION SCIENCE	22
PHYSICAL REHABILITATION SCIENCE	15
NURSING	15
LAW / CRIMINOLOGY	13
INFO. TECH. / APPLIED SCIENCES	9
OTHER HEALTH SCIENCE	91
OTHER SOCIAL SCIENCE	90

Overview

1. What is a Complex Sample?
   a) Random Stratification
   b) Clustering
   c) Multistage
2. Weighting of Cases
   a) Appropriate for only Proportions and Means
3. Bootstrap Estimation
   a) Variances and related statistics (standard errors, correlations, t-tests, etc.)
   b) Appropriate for most statistical analyses

1. Complex Samples (Surveys)

- **Standard** statistical procedures taught in textbooks and available in most statistical software packages (SPSS, R, SAS, Stata, etc.) assume that the data is based upon a ***random sample*** of the population
- Most psychological research, however, employs ***samples of convenience***
- Although random samples provide the most accurate representation of the population, they are not very efficient
- Statistics Canada employs ***complex samples*** because they are more efficient (requiring smaller samples for the same level of precision as larger random samples)
- Analyses of complex samples require **advanced** statistical procedures

What is a Complex Sample?

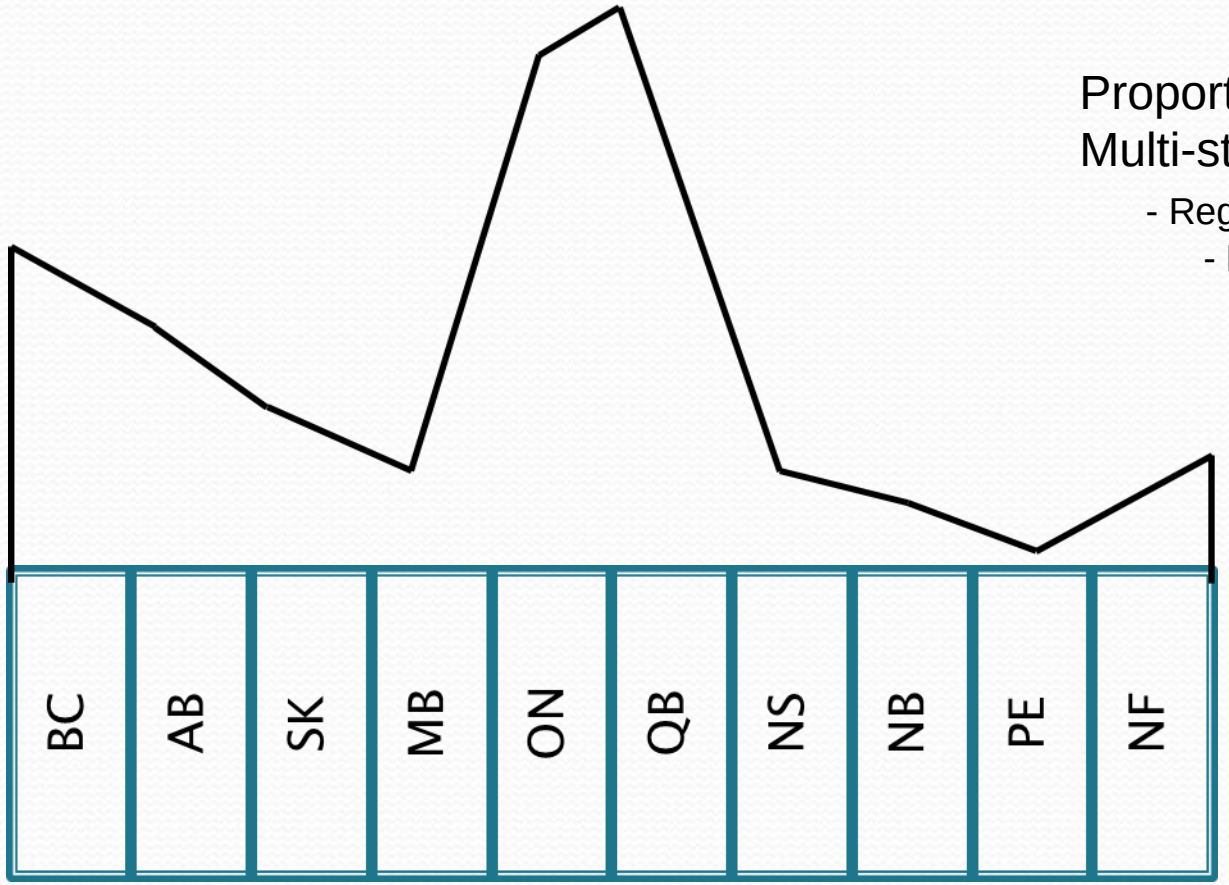
- ***Complex sampling*** is a generic term usually referring to a combination of several different advanced sampling techniques
- Statistics Canada employs multistage random stratification cluster sampling
  - Multistage
    - Nests the other sampling techniques in a hierarchical manner
  - Random Stratification
    - Stratifies the population into subgroups (called ***strata***) based upon key demographic variables, and then randomly draws a sample from each
  - Clustering
    - More efficient data-collection for a geographically diverse country

Random vs. Random Stratified Sampling?

Multistage Randomized Stratified Cluster Sampling (CCHS, 2012)

- Desired representative samples for age & sex (Stratification)
  - Target 27,500 sample size (constrained by budget)
  - 4 age groups (15-24, 25-44, 45-64, 65+)
- 8 groups (age & sex) × 125 cases × 10 Provinces
  - 10,000 cases + 17,500 cases proportional to population size
- Over-sample 43,030 to account for nonresponses
- Stage 1: Clusters (geographical regions based on Labour Force Survey sampling frame) ***assuming that neighbours are similar to each other***
  - Stage 2: Dwellings or household within each cluster
    - Stage 3: Randomly select a member of each household
- Yields a BIASED sample (will be readjusted latter)

Sampling (Still problem with nonresponse)



Proportional to Population Size
Multi-stage clusters:

- Region
  - Dwelling
    - Person

N = 10 x 1000

Stratified Random
(Sex X Age Group)

2. Weighting Cases (Fix Sex Breakdown)

Unweighted (Sample) Data

```
. 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
```

Weighted (Population) Data

```
. 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
```

Using Public Use Microdata File (PUMF) version of Canadian Community Health Survey (CCHS, 2018)

Provincial/Territorial Counts

Unweighted (Sample) Data

```
. 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
```

Weighted (Population) Data

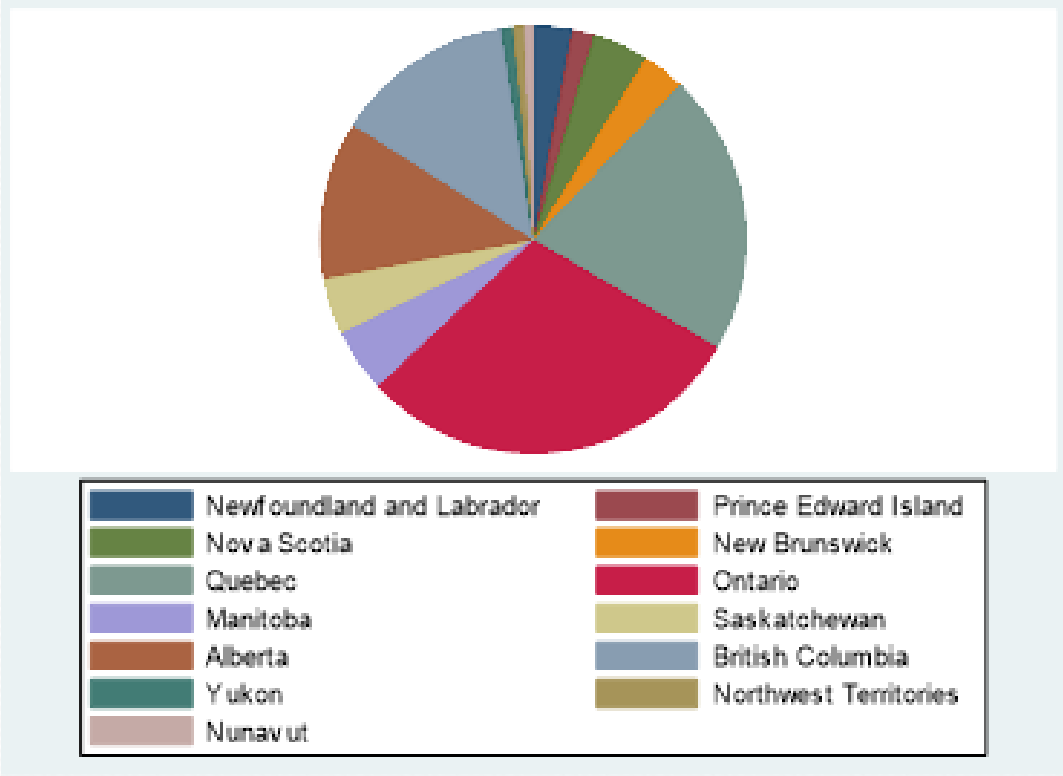
```
. tabulate GEO_PRV [iweight = WTS_M]

  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
```

Using random sampling we would require 770 / .0009 = 855,556 cases in sample to get 770 cases in Nunavut

Provincial/Territorial Counts Pt. 2

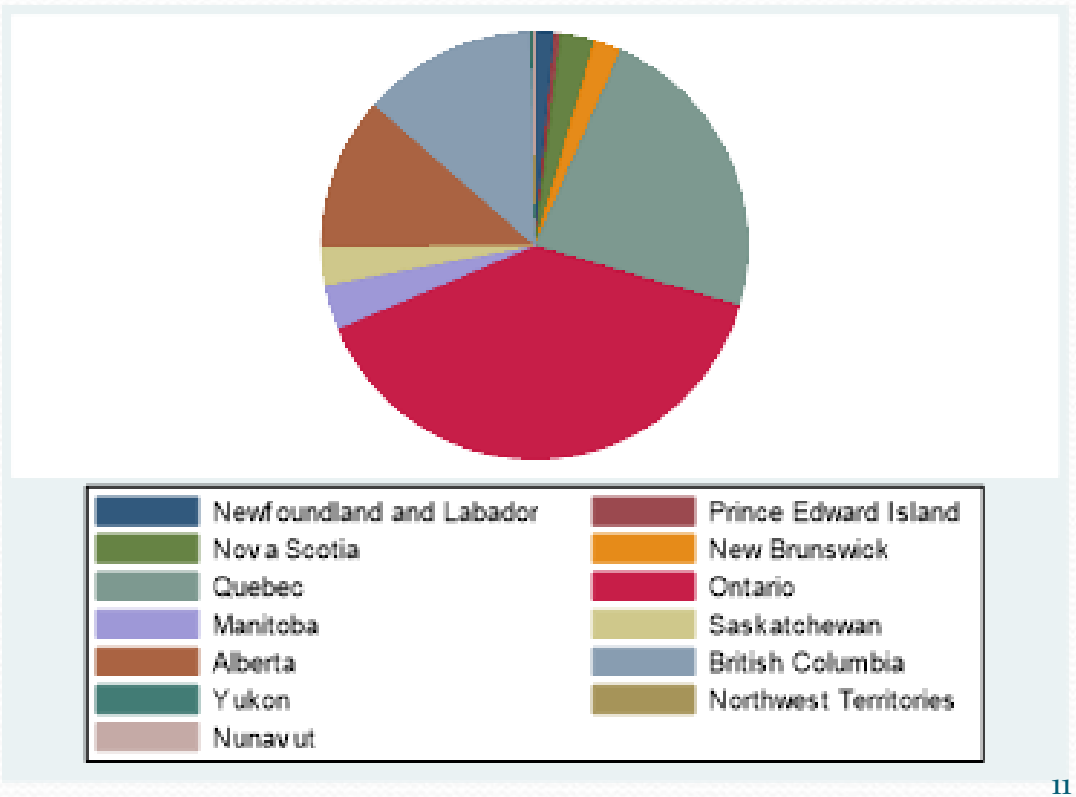
Unweighted (Sample) Pie Chart

- `. graph pie, over(GEO_PRV)`



Weighted (Population) Pie Chart

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



3. Bootstrapping (***Slow** food*)

- Weighting of cases is okay for estimating population means and proportions
- However, weighting makes a ***mess*** of estimating population variances & related statistics (like standard errors, t-tests, correlations, etc.)
- Bootstrap estimation is extensive re-sampling of statistical estimates from a given sample (with replacement of a case after each draw)
- Statistics Canada employs bootstrap estimation of variance (and all related statistics) using 500 or 1000 samples
- Distribution of underlying statistic is empirically derived vs. being based upon a theoretical distribution (with required assumptions being met)

Standard vs. Bootstrap Statistical Analysis

Standard Procedure

- Distribution of raw scores
  - Sample mean is good estimate of population mean $\bar{X} \cong \mu$
  - SD can be used to estimate SE_{mean}

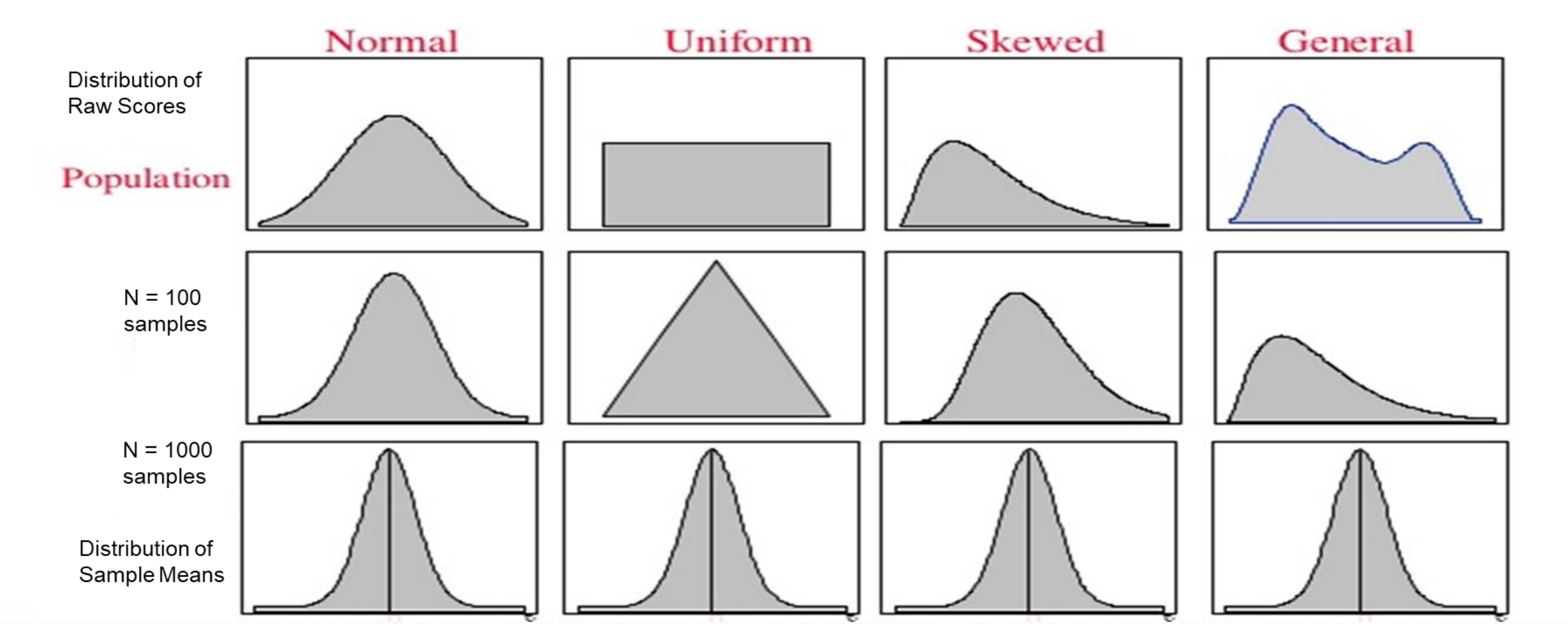
$$SE_{mean} = \frac{SD}{\sqrt{N}}$$

  - SE used in significance tests, 95% CI, etc.

Bootstrap Procedure

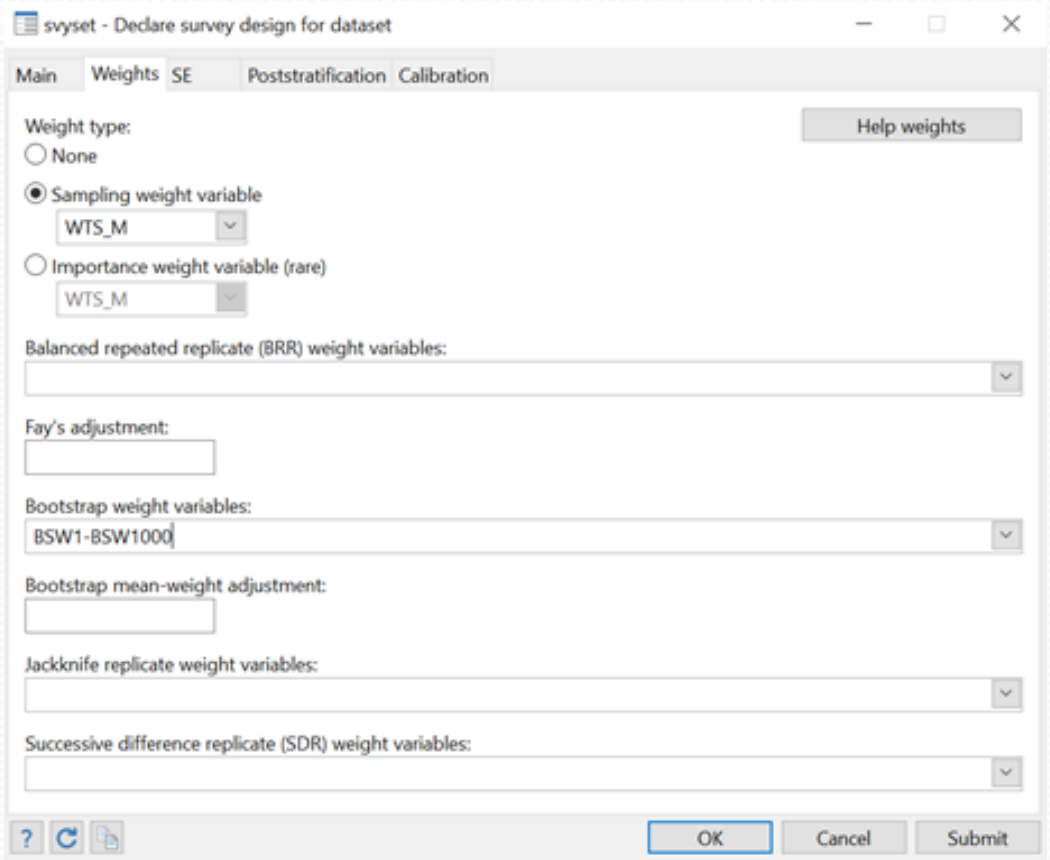
- Empirical distribution of means from 1000 bootstrap samples
  - "Mean of 1000 sample means" is a better estimate of population mean
  - $SE_{mean} = SD_{means}$
  - Standard error (SE) = sampling distribution of statistic in population

Central Limit Theorem (Bootstrapping)

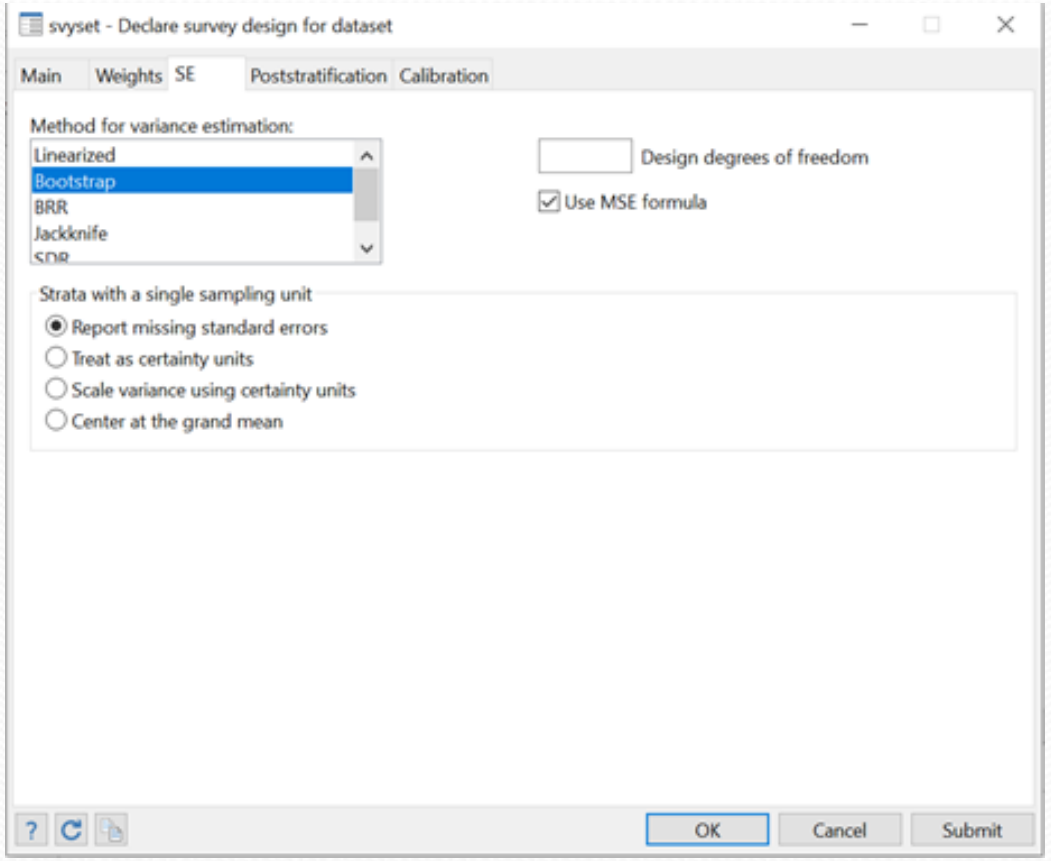


Setting Up Survey Design in Stata (SVYSET)

- Weighting Tab



- Standard Error (SE) Tab



a) Estimating Proportions with Bootstrapping

- **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>
```

- **Stata syntax for TABULATE command with bootstrapping:**

`. 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
```

b) Estimating a t-test with Bootstrapping

```
. svy : mean hwtdghtm, over(DHH_SEX) coeflegend

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

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]
------------------------------------------------------------------------

. 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
------------------------------------------------------------------------------
```

c) Estimating a correlation with Bootstrapping

```
. svy : regress hwtdghtm DHH_SEX

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
```

```
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
------------------------------------------------------------------------------
```

Find the correlation (same sign of coefficient)

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

Find Z^2 or Wald's χ^2 or F

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

Complex Survey Procedures in Major Statistical Packages

SPSS	SAS	Stata	R
Weighting of Cases: Means Proportions *(**Note:** SPSS's approach to analyzing complex samples is different from Statistics Canada)*	PROC - SURVEYSELECT SURVEYFREQ SURVEYMEANS SURVEYLOGISTIC SURVEYREG	Setup and utilities Tables Means, proportions, ratios, totals Linear models Binary outcomes Ordinal outcomes Categorical outcomes Count outcomes Fractional outcomes Survival models Multilevel mixed-effects models Endogenous covariates Sample-selection models Generalized linear model (GLM) SEM (structural equation models) LCA (latent class analysis) FMM (finite mixture models) IRT (item response theory) DEFF, MEFF and other statistics Resampling	svrvar svyby svycdf svychisq svyciprop svycontrast svycoplot svycoxph svyCprod svycralpha svydesign svyfactanal svyglm svyhist svyivreg svykappa svykm svyloglin svylogrank svymean svymle svynls svyolr svyplot svyprcomp svypredmeans svyquantile svyranktest svyratio svyrecvar svysmooth svystandardize svysurvreg svytable svytotal svyttest

Contact Information

- richard.maclennan@uregina.ca
- regina.rdc@uregina.ca
- www.uregina.ca/research/rdc

University of Regina

Government of Saskatchewan

CRDCN RCCDR

Statistics Canada Statistique Canada

Canada

Contact Us

Academic Director:

Dr. Richard Maclennan

Analyst:

Dr. Jennifer McConnell-Nzunga

Innovation Place, Room 190

#2 Research Drive

Open Monday to Friday 8:30am - 4:30pm

T: 306-585-1122

regina.rdc@uregina.ca

www.uregina.ca/research/rdc