Survey Sampling with IBM SPSS Statistics

Mārtiņš Liberts

Central Statistical Bureau of Latvia

16–20 February 2014

Content

Content

IBM SPSS Statistics

- SPSS Statistics is not a magic box
- SPSS Statistics is a tool – it has to be used correctly
- Keep an eye on alternatives (R) – do not limit your tool set

SPSS Statistics Base vs SPSS Complex Samples

- Statistics Base
  - (+) Flexibility
  - (+) Full control of the sampling process
  - (–) More coding is necessary
- Complex Samples
  - (+) Easy to implement
  - (–) Less flexibility
  - (–) Algorithms have to be know (do not use as a black box)
  - (–) Extra coding could be needed

GUI vs Syntax

- Graphical user interface (GUI)
  - (+) Good for trying, exploring and learning
  - (+) Good for small tasks to get fast results (table of frequencies)
  - (–) Not good for production
  - (–) Hard to reproduce
  - (–) Hard to keep a track

GUI vs Syntax

- Syntax
  - (+) Good for complex tasks (data management)
  - (+) Good for repetition
  - (+) Some features are available only through syntax
  - (+) Must be used for production
  - (+) Reproducible, easy to keep a track
  - (–) You have to adapt to coding (if not done before)

Using Syntax

- Write all procedures as syntax – syntax should represent whole process
- Use 'Paste' instead of 'OK'

Help

- SPSS Help
- SPSS Command Syntax Reference (must have)
- SPSS mailing list (`http://spssx-discussion.1045642.n5.nabble.com`)
- SPSS Tools by Raynald Levesque (`http://spsstools.net`)
- StackOverflow (`http://stackoverflow.com/questions/tagged/spss`)
- Google

Content

Introduction

Sampling

Three examples

Sampling examples

Estimation examples

Random Generators

There are two pseudo random generators:

- **MC**
  - used in version 12 and previous releases
  - `set rng = mc.`
- **MT**
  - Mersenne Twister random number generator
  - `set rng = mt.`
- `show rng.`

Random Generators

- Make the sampling process reproducible
- Set the initialization value before drawing a final sample
- **MC**
  - `set seed = 987654.`
- **MT**
  - `set mtindex = 123456.`

Content

Sampling topics covered

- Sampling
- Design weighting
- Estimation of population parameters
- Estimation of precision

Sampling designs

- Simple random sampling (SRS)
- Stratified simple random sampling (SSRS)
- Two stage sampling:
    1. Stratified systematic πps sampling
    2. SRS

Artificial data set

- Represents artificial household population
- $N_{PSU} = 200$
- $N_H = 50227$
- $N_P = 109371$
- Same procedures can be used also for economic and agriculture populations

Content

SRS

- $n = 1000$
- $w_i = \frac{N}{n}$

SSRS

- $n = \{800, 150, 50\}$
- $w_{hi} = \frac{N_h}{n_h}$

Two stage

- $n = 1000$
  1. $H = 4$ strata
  2. $n_h = 10$ PSUs in each strata
  3. $l_h = 25$ households in each sampled PSU
- Sampling step $s_h = \frac{M_h}{n_h}$
- $w_{hi} = \frac{M_h}{n_h M_{hi}} \frac{M_{hi}}{l_h}$

Content

Estimates

$$\hat{\Theta} = \sum_{i=1}^{n} y_i w_i$$

$$v\left(\hat{\Theta}\right)$$

$$se\left(\hat{\Theta}\right) = \sqrt{v\left(\hat{\Theta}\right)}$$

$$cv\left(\hat{\Theta}\right) = \frac{se\left(\hat{\Theta}\right)}{\hat{\Theta}}$$

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{v_{SRS}\left(\hat{\Theta}\right)}$$

SRS

$$v\left(\hat{\Theta}\right) = N^2 \frac{1-f}{n} s^2$$

$$f = \frac{n}{N}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

SSRS

$$v\left(\hat{\Theta}\right) = \sum_{h=1}^{H} N_h^2 \frac{1-f_h}{n_h} s_h^2$$

$$f_h = \frac{n_h}{N_h}$$

$$s_h^2 = \frac{1}{n_h-1} \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2$$

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$$

Two Stage

Variance approximation – **Ultimate cluster method** (Hansen M., Hurwitz W. and Madow W. (1953). *Sample Survey Methods and Theory*. New York: Wiley)

$v_1\left(\hat{\Theta}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(z_{hi\cdot} - \bar{z}_{h\cdot\cdot}\right)^2$

$z_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij} z_{hij}$

$\bar{z}_{h\cdot\cdot} = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi\cdot}$

Two Stage

Variance approximation – **Ultimate cluster method** (Hansen M., Hurwitz W. and Madow W. (1953). *Sample Survey Methods and Theory*. New York: Wiley)

$$v_1\left(\hat{\Theta}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(z_{hi\cdot} - \bar{z}_{h\cdot\cdot}\right)^2$$

$$z_{hi\cdot} = \sum_{j=1}^{m_{hi}} w_{hij} z_{hij}$$

$$\bar{z}_{h\cdot\cdot} = \frac{1}{n_h} \sum_{i=1}^{n_h} z_{hi\cdot}$$

$$v_2\left(\hat{\Theta}\right) = (1 - f_h) \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(z_{hi\cdot} - \bar{z}_{h\cdot\cdot}\right)^2$$

Design Effect

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{v_{SRS}\left(\hat{\Theta}\right)}$$

Design Effect

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{v_{SRS}\left(\hat{\Theta}\right)}$$

$$V_{SRS}\left(\hat{\Theta}\right) = N^2 \frac{1-f}{n} S^2$$

Design Effect

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{v_{SRS}\left(\hat{\Theta}\right)}$$

$$V_{SRS}\left(\hat{\Theta}\right) = N^2 \frac{1-f}{n} S^2$$

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2 = \frac{1}{N-1} \sum_{i=1}^{N} y_i^2 - \frac{N}{N-1} \bar{Y}^2$$

Design Effect

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{v_{SRS}\left(\hat{\Theta}\right)}$$

$$V_{SRS}\left(\hat{\Theta}\right) = N^2\frac{1-f}{n}S^2$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N} y_i^2 - \frac{N}{N-1}\bar{Y}^2$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{n} y_i^2 w_i - \frac{1}{N(N-1)}\left(\sum_{i=1}^{n} y_i w_i\right)^2$$

Design Effect

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{v_{SRS}\left(\hat{\Theta}\right)}$$

$$V_{SRS}\left(\hat{\Theta}\right) = N^2\frac{1-f}{n}S^2$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N} y_i^2 - \frac{N}{N-1}\bar{Y}^2$$

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{n} y_i^2 w_i - \frac{1}{N(N-1)}\left(\sum_{i=1}^{n} y_i w_i\right)^2$$

$$deff\left(\hat{\Theta}\right) = \frac{v\left(\hat{\Theta}\right)}{N^2\frac{1-f}{n}\left(\frac{1}{N-1}\sum_{i=1}^{n} y_i^2 w_i - \frac{1}{N(N-1)}\left(\sum_{i=1}^{n} y_i w_i\right)^2\right)}$$

Thank you!