Sample Size Calculation

Mārtiņš Liberts

Central Statistical Bureau of Latvia

16–20 February 2014

Content

Content

Introduction

- Sample size calculation with an aim to meet **survey requirements** is probably the hardest task in survey sampling
- Survey requirements:
  - To estimate the key population parameters with required precision
  - To estimate many other population parameters

It is hard because

- There is not enough information about the target population (especially at unit level)
- Assumptions have to be made
- Survey requirements can be contradictory
- The achieved precision depend on many non-sampling aspects:
  - Non-response
  - Frame errors
  - Measurement errors
- The survey budget is limited

Content

Introduction

- Calculation of sample size is easier if population parameter is proportion

$$p = \frac{1}{N} \sum_{U} y_i$$

$$y_i \in \{0, 1\}$$

- Assumption: population size N is known

SRS

$$\hat{p} = \frac{1}{N} \sum_{s} y_i w_i$$

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

$$\hat{p} = \frac{1}{N} \frac{N}{n} \sum_{s} y_i = \frac{1}{n} \sum_{s} y_i$$

SRS

$$V\left(\hat{p}\right) = \frac{1 - \frac{n}{N}}{n} S^2$$

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

$$S^2 = \frac{N}{N-1} p \left(1 - p\right)$$

$$V\left(\hat{p}\right) = \frac{1 - \frac{n}{N}}{n} \frac{N}{N-1} p \left(1 - p\right)$$

SRS

$$SE\left(\hat{p}\right) = \sqrt{\frac{1-\frac{n}{N}}{n}\frac{N}{N-1}p\left(1-p\right)}$$

$$MoE\left(\hat{p}\right) = Z_{\beta}\sqrt{\frac{1-\frac{n}{N}}{n}\frac{N}{N-1}p\left(1-p\right)}$$

$$Z_{.95} = 1.96$$

For example:

- If $MoE = .01$, the confidence interval is $\pm .01$.
- We can conclude that $|p - \hat{p}| < .01$ with probability .95.

The question to be asked to the user of statistics:

- What is **the largest acceptable error**?
- What is **the acceptable margin of error**?

SRS

$$MoE\left(\hat{p}\right) = Z_{\beta}\sqrt{\frac{1-\frac{n}{N}}{n}\frac{N}{N-1}p\left(1-p\right)}$$

$$\Downarrow$$

$$n = \frac{N}{\frac{\left(MoE(\hat{p})\right)^{2}(N-1)}{Z_{\alpha}^{2}p(1-p)}+1}$$

$$\max_{p}\left(n\right) = \frac{N}{\frac{\left(MoE(\hat{p})\right)^{2}(N-1)}{.25Z_{\alpha}^{2}}+1}$$

SSRS

If stratification is done by the domains of interest

$$n_d = \frac{N_d}{\frac{(MoE(\hat{p}))^2 (N_d - 1)}{Z_\alpha^2 p(1-p)} + 1}$$

In case of non-response

$$m_d = \min\left(N_d; \left\lfloor \frac{n_d}{r} \right\rfloor + 1\right)$$

where r is a response rate

Sample Size Calculator

- Excel file `Sample_size_calc_ver021_ENG.xlsx`
- `http://www.surveysystem.com/sscalc.htm`

Content

Variance with two sample sizes

$$V_0 = N^2 \frac{S^2}{n_0}$$
$$V_1 = N^2 \frac{S^2}{n_1}$$

$$\Downarrow$$

$$\frac{V_0}{V_1} = \frac{n_1}{n_0}$$

Variance with two sample sizes

$$n_1 = \frac{V_0 n_0}{V_1}$$

If $\frac{V_1}{V_0} = \frac{1}{2} \Rightarrow n_1 = 2n_0$

SE and CV with two sample sizes

$$\frac{SE_0^2}{SE_1^2} = \frac{n_1}{n_0}$$

$$\frac{CV_0^2}{CV_1^2} = \frac{n_1}{n_0}$$

$$n_1 = \frac{SE_0^2 n_0}{SE_1^2} = \frac{CV_0^2 n_0}{CV_1^2}$$

If $\frac{SE_1}{SE_0} = \frac{1}{2} \Rightarrow \frac{SE_1^2}{SE_0^2} = \frac{1}{4} \Rightarrow n_1 = 4n_0$

SE and CV with two sample sizes

Example: $CV_0 = 0.05$ and $CV_1 = 0.04$

$$\frac{CV_1}{CV_0} = \frac{4}{5} = 0.8$$

$$n_1 = \frac{n_0}{0.8^2} = \frac{n_0}{0.64} = 1.5625 n_0$$

Content

Task

- Task: to calculate a sample size and sample allocation for stratified sampling design with an aim to achieve specified precision for domains.
- Conditions:
  - The domains of interest are non-overlapping
  - Each stratum belongs to only one domain
  - Populations size N_{dh} is available for each stratum
  - Populations total Y_{dh} is available for each stratum
  - Estimates of S^2_{dh} are available for each stratum

The Population

Domain	Str1	Str2	Str3	Str4	Str5	Str6	Str7	Str8
Dom1	N_{11}	N_{12}	N_{13}	N_{14}	N_{15}	N_{16}	N_{17}	N_{18}
Dom2	N_{21}	N_{22}	N_{23}	N_{24}	N_{25}	N_{26}	N_{27}	N_{28}
Dom3	N_{31}	N_{32}	N_{33}	N_{34}	N_{35}	N_{36}	N_{37}	N_{38}
Dom4	N_{41}	N_{42}	N_{43}	N_{44}	N_{45}	N_{46}	N_{47}	N_{48}

The Population

Domain	Str1	Str2	Str3	Str4	Str5	Str6	Str7	Str8
Dom1	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{16}	Y_{17}	Y_{18}
Dom2	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}	Y_{26}	Y_{27}	Y_{28}
Dom3	Y_{31}	Y_{32}	Y_{33}	Y_{34}	Y_{35}	Y_{36}	Y_{37}	Y_{38}
Dom4	Y_{41}	Y_{42}	Y_{43}	Y_{44}	Y_{45}	Y_{46}	Y_{47}	Y_{48}

The Population

Domain	Str1	Str2	Str3	Str4	Str5	Str6	Str7	Str8
Dom1	S^2_{11}	S^2_{12}	S^2_{13}	S^2_{14}	S^2_{15}	S^2_{16}	S^2_{17}	S^2_{18}
Dom2	S^2_{21}	S^2_{22}	S^2_{23}	S^2_{24}	S^2_{25}	S^2_{26}	S^2_{27}	S^2_{28}
Dom3	S^2_{31}	S^2_{32}	S^2_{33}	S^2_{34}	S^2_{35}	S^2_{36}	S^2_{37}	S^2_{38}
Dom4	S^2_{41}	S^2_{42}	S^2_{43}	S^2_{44}	S^2_{45}	S^2_{46}	S^2_{47}	S^2_{48}

The Procedure

- Set $n_d = n_0$
- Compute Neyman allocation for domain d as:

$$n_{dh} = n_d \frac{N_{dh} S_{dh}}{\sum_{h=1}^{H} N_{dh} S_{dh}}$$

- Compute expected variance for the estimate of total in domain d as:

$$V\left(\hat{Y}_d\right) = \sum_{h=1}^{H} N_{dh}^2 \frac{\left(1 - \frac{n_{dh}}{N_{dh}}\right)}{n_{dh}} S_{dh}^2$$

The Procedure

- Compute expected CV for the estimate of total in domain d as:

$$CV\left(\hat{Y}_d\right) = \frac{\sqrt{V\left(\hat{Y}_d\right)}}{Y_d}$$

- Compare expected CV with required CV
- if $CV\left(\hat{Y}_d\right) > CV^*$: increase sample size for domain d: $n_d := n_d + 1$
- if $CV\left(\hat{Y}_d\right) \leq CV^*$: n_d is the final sample size for domain d.

The Procedure

- Iterative procedure
- R code

Content

Resume

- Sample size calculation for SRS if the population parameter is a proportion
- Relationship between sample size, variance and coefficient of variation
- Sample size and sample allocation calculation for SSRS

Conclusions

- The proposed solutions work only in specific cases
- In other cases they may work but may fail as well
- The solutions can be used as a leading information with caution

Thank you!