PRASC

Project for the Regional Advancement of Statistics in the Caribbean

Projet régional pour l'avancement de la statistique dans les Caraïbes

In partnership with Canada

Methodological Considerations for Frame and Sample Design

Project for the Regional Advancement of Statistics in the Caribbean (PRASC)

Delivering insight through data for a better Canada

Anne-Marie Côté
November 23, 2022, Barbados

Analytical Question

The government wants to design an initiative to support the **accommodation and food services industry** in your country to recover from the effect that the global pandemic had on that industry.

The Ministry of Finance has the responsibility for the economic development of this industry and wants to gather information in order to make guided decisions on how the initiative should be conducted.

What do we need to estimate?

- Who should we target?
- What are our variable of interest?
- Which variables are most important?
- Domains of interest
- Are there auxiliary data available which could be used to improve the sample plan?
  - Are there auxiliary data which are related to the variable(s) of interest?

Steps of our sample design

1. Population
2. Frame
3. Survey units
4. Sample size
5. Stratification /Sample allocation
6. Sample selection method

Using the SBR as a Frame

1. Extract the SUF if not done systematically on a regular basis by your organisation
2. Work from a copy!
3. Verification of the information on the frame

Using the SBR as a Frame

- Clean the SUF by only keeping the records that are in your population of interest
  - Alive
  - Business Type
  - Country of control / Legal country
  - Establishment level
  - Prospect Flag
  - ISIC of interest
  - Duplicates ?

Using the SBR as a Frame

- Basic verification of the coverage of the frame
  - Using subject-matter expertise, ensure that:
    - The large units are listed (in general, compared to last cycle, per geo/industry)
    - The must take units are there (large and odd ones)
    - Large units based on the auxiliary variable
  - Comparison from one cycle to the next
    - Births and deaths
    - # of units in each geo/industry
    - Units that change geo/industry from one cycle to the next
  - Compare with other information (if available) to verify the coverage

Using the SBR as a Frame

- Verification of the auxiliary variable
  - Do we have a value for all of the units?
    - If not, is it just for a few units or a large number?
      - Large number, is there another variable that is correlated for which we have for everyone?
      - A few, how can we impute it?

Sample Design

- Identification of must-take units
- Stratification
  - Geo and industry
  - Size variable
- Method: Gunning and Horgan

Gunning and Horgan

- Step 1: We order the data in ascending order by IndustryGroupDescriptio, MusTTakeFlag and Design variable
- Step 2: We define the must-take stratum for each IndustryGroupDescription when the MustTakeFlag =1.
- Step 3: We determine the minimun and maximun value for each IndustryGroupDescription when MustTakeFlag = 0.
- Step 4: In our case, we want 3 strata (including the MT) for each IndustryGroupDescription. So 2 strata need to be created (Take None and Take Few). To do that we apply the Gunning and Horgan method. For us, L=2

Gunning and Horgan

$$b_h = \min(x) * \left\lfloor \frac{\max(x)}{\min(x)} \right\rfloor^{\frac{h}{L}} \quad for\ h = 1,2,\ldots,L-1$$

Where

X -> Auxiliary variable (Design variable)

L -> Number of groups / strata to be created

h -> Stratum identifier

b -> Boundary

Gunning and Horgan

- For example, IndustryGroupDescription=BRB_Accom and MustTakeFlag=0

- b1 = 256 * (9,448,064 / 256)$^{1/2}$
- b1 = 49180.33

- So, we put BRB_Accom_TF (Take Few) if the value of the Design variable is between [0, 49,180.33] and we put BRB_Accom_TS (Take Some) for the rest.

Sample Size

- Determine the sample size based on the budget, the quality of the estimates (expected CV) and the response rate.
- Calculate the sample size for each stratum by using the Root-N allocation method

$$n_h = n \frac{\sqrt{N_h}}{\sum \sqrt{N_h}}$$

Root-N allocation method

$$n_h = n \frac{\sqrt{N_h}}{\sum \sqrt{N_h}}$$

- At this step, we are assuming a 100% response rate
- Apply a minimum sample size in each stratum (like 5)

$$If\ N_h \leq 5\ then\ n_{h\,final} = N_h$$

$$\text{else if } N_h > 5\ and\ n_h < 5\ then\ n_{h\,final} = 5$$

$$\text{else } n_{h\,final} = n_h$$

Expected CV

- Calculate of the expected CV to see what they will be with a response rate of 100%
- Adjust the sample size of some strata if needed based on the expected CV

Nonresponse adjustment

- Since expected CVs were calculated considering a response rate of 100%, increase the sample size to account for nonresponse.
- Adjust the sample size inside each stratum for nonresponse rate. We can use an overall nonresponse rate or a stratum specific nonresponse rate.

$$nh\,adjusted = Min\left(round\left(\frac{nh}{(1-NonResponseRate)}\right), Nh\right)$$

Sample Selection

- Generic PRASC SBR application includes a Permanent Random Number (PRN)
  - A random number between 0 and 1
  - Assigned when a unit is added to the SBR, and is retained throughout all SBR updates
    - Ex: If inactivated then reactivated, same PRN is retained
  - Two ways to use:
    - Percentages – if want a 10% sample select an interval of 0.1 (ex. 0.20 to 0.30)
      - Will not always produce the exact number of allocated units, especially in small populations
    - Counting – select a starting point and count number of sampled units
      - A bit more cumbersome to select the sample

Permanent Random Number (PRN)

- Advantages of a PRN
  - New and inactivated units automatically accounted for
  - Allows easy sample rotation (control of response burden)
  - Produces a simple random sample which simplifies unbiased estimation

You can contact the PRASC team at:

statcan.prasc-prasc.statcan@canada.ca