Business Surveys: Methodological Considerations for Frame, Sample Design and Estimation

Last updated: June 8, 2023

Frame

1. Purpose of the frame

- Defines the population of interest
- Should provide the means of identifying and contacting the units of the survey population
- Ideally, it should contain auxiliary variables that can be used in the survey

2. Information/tool that we know of

- The Statistical Business Register (SBR) containing all records and variables, including the Permanent Random Number (PRN).

3. What that means in practice

- Use the SBR to create a Survey Universe File (SUF) file that will be used as the initial frame for the survey.
- Add to the SBR or the SUF file the IndustryGroup (i.e., grouped by International Standard Industrial Classification (ISIC)) that will be used in the sampling strategy.

- Clean the SUF file by only keeping the records in your population of interest because the estimates will reflect the frame you are using.
    - This is done by removing some records you don't want in your final frame. For example,
        - based on the ISIC code, for example, IndustryGroup in (130) or
        - ProspectiveLegalFlag=0

Sample Design

1. Purpose of the sample design

- To select a random subset of units to draw inferences for the whole survey population.

2. Information/tools that we know of

- Normally, the sample design is established based on (among other things):
    - The population of interest (size, availability of the data, etc.)
    - Cost/timeliness
    - Precision
    - Domains of interest
- Distinctive feature of business surveys: The business populations are generally skewed.
    - A small number of very large businesses account for the majority of the activity

- Within each domain of interest (such as industry group), the population is usually split into strata based on an auxiliary variable.
    - Must-take units (large units sampled with certainty)
    - One or two take-some portions (smaller units from which a random sample is drawn)

3. What that means in practice

Must-take units

- The must-take units are surveyed with certainty (probability of one). They are the ones driving your estimates (big units in each IndustryGroup).
- Creation of a new flag, MustTakeFlag, in the frame indicating which records are Must-Take.

Stratification variables

- The first goal of the stratification variable is to have homogeneous strata, i.e., within a given stratum, the units are similar. Normally, the strata are based on the subgroups of the population for which you are planning to release your survey estimates. These are called estimation domains. For example, in your situation, we have ISIC publication groups.
- Add a size variable based on an auxiliary quantitative variable (for example, revenue or wages). This auxiliary variable needs to be present for all units on the frame. The higher the correlation with the survey's principal variable of interest, the better.
- In your case, the size variable, wages, is used by the Gunning and Horgan Method to define strata boundaries.

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

Where

X ➔ Auxiliary variable (wages)
L ➔ Number of groups / strata to be created
h ➔ Stratum identifier
b ➔ Boundary

- The must-take units are excluded from the boundary calculation because there are selected with certainty and must be put in an extra category.
- You will find below a document explaining the Gunning and Horgan Method:

PRASC_Reference_Stratification - Grouping

- You will also find an example at the end of the document.

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

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

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

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

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

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

  - See the example at the end of the document.
- You will find below a document explaining the Root-N allocation Method

PRASC_Reference_RootN Allocation_JUL20

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
- See the example at the end of the document.

Nonresponse rate 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. You can use an overall nonresponse rate or a stratum specific nonresponse rate
  - $n_h\ adjusted = Min\ (round\left(\frac{n_h}{(1-NonResponseRate)}\right), N_h)$
  - See the example at the end of the document.

Sample selection for the first cycle[1]

- Use the Permanent Random Numbers (PRNs) method to select a simple random sample from each stratum; a PRN takes on a value between 0 and 1.
  - The SBR automatically generates a PRN for each business; this remains constant over time.
  - It is essential that the PRN associated to one unit never changes (ensure appropriate quality assurance practices when manipulating data files during processing).
- Within each stratum, select the exact sample size[2]
  - Sort the units by ascending PRN
  - Choose a starting point (for example, 0)
  - Sequentially select the desired number of units
- See the example at the end of the document.
- You will find below a document explaining in more details the PRN method. We will talk at a later time about birth, death, stratum jumper, sample rotation, coordinated sampling and weighting.

PRASC_Reference_Using PRN to Select and

[1] This section is a very quick summary of the document written by Kim Fyfe, "Using PRN to Select and Maintain Survey Samples," that you will find at the end of this section.

[2] See the "Using PRN to Select and Maintain Survey Samples" document for further details on an alternative method for using the PRN to select a sample.

Estimation

1. Purpose of estimation

- A way to obtain values for the population of interest and to draw conclusions about this population based on the information obtained from only a sample of the population.

2. Information/tools that we know of

- The principle behind estimation in a probability survey is that each sample unit represents not only itself, but also several units of the survey population. We call it the design weight.
- Determining this design weight is an important part of the estimation process. The design weight can be adjusted to account for nonresponse.
- Once the final estimation weights have been calculated, they are applied to the sample data in order to compute estimates.
- An important part of estimation is estimating the magnitude of the sampling error in the estimate. This provides a measure of the quality of the survey's estimates for the specific sample design.

3. What that means in practice

- Under the sampling design assumptions, we calculate an estimator of the population total and a variance estimator.
- In our case, the sampling design is a stratified random sampling without replacement.
- A stratum jumper arises when the stratification information collected in the field is different from the information in the sampling frame (for example the industryGroup). These differences can be explained by errors on the sampling frame, which are partly due to outdated information, and the time lag between sampling and collection.

The case of 100% response

- An estimator of the domain total with H strata that cut across the D domains

$$\widehat{Total}_d = \sum_h \frac{N_h}{n_h} \sum_{s_h} \breve{y}_k$$

where

- $\frac{N_h}{n_h}$ *is the design weight in stratum h*
- $\breve{y}_k = y_k$ *if k is in the domain* U_d *and 0 otherwise*

- An estimator of the domain variance with H strata that cut across the D domains

$$\widehat{Variance}_d = \sum_h N_h^2 \frac{1 - f_h}{n_h} S^2_{\breve{y}s_h}$$

where

- $f_h = \frac{n_h}{N_h}$ *is the sampling fraction in stratum h*
- $S^2_{\breve{y}s_h} = \frac{1}{(n_h - 1)} \sum_{s_h} (\breve{y}_k - \bar{\breve{y}}_{s_h})^2$
- $\bar{\breve{y}}_{s_h} = \sum_{s_h} \breve{y}_k / n_h$

- $\check{y}_k = y_k$ *if k is in the domain* U_d *and 0 otherwise*

- See the example at the end of the document.

The case of nonresponse

- All surveys are affected by nonresponse. We can categorize nonresponse in two main categories:

Partial nonresponse

- This happens when we have missing information for some variables or a complete section of the questionnaire is missing for some sampled units.
- In this case, we will impute the missing values by using an imputation method.

Total nonresponse

- This occurs when all or almost all information for a sampled unit is missing. For example, the unit does not want to participate to the survey or we are not able to contact the unit or the information we received is unusable.
- There are two ways to handle total nonresponse in a survey:
    - Imputation approach
    - Reweighting approach

Imputation approach

- We impute the missing information using an imputation method.
- Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data. This is done by changing some of the responses and all of the missing values on the unit being edited to ensure that a plausible, internally consistent unit is created.
- Imputation has the advantage of leading to complete the dataset so no information is lost and can improve the quality of the final data. On the downside, it is possible to forget that this file contains a mix of real data and imputed data. Also, imputed data could have an impact on final estimates. After imputation, the dataset should normally only contain plausible and internally consistent data that can then be used for estimation of the population of interest. They are many imputation methods available but you need to keep in mind that the overall guiding should be simplicity.
- When it's time to select imputation method:
    - Determine and evaluate the possibility of using auxiliary variables to perform the imputation in order to approximate as accurately as possible the unknown missing values and thus to produce quality estimates of the population of interest. An auxiliary variables need to be available for units with and without variables to be imputed and they should be related as much as possible to the variables to be imputed in order to reduce nonresponse bias and variance;
    - Imputation method should be chosen carefully, considering the type of data to be imputed;
    - Imputation method should aim to reduce the nonresponse bias and preserve relationships between questions as much as possible.
- When we perform imputation:
    - Should impute the minimum number of variables to preserving as much respondent data as possible;

- Imputed units should satisfy all edits;
- Determine imputation classes because imputation is usually performed within subgroups of the population (e.g., stratum, industry group). The imputation classes can be hierarchical. We start the imputation with very detailed classes (e.g., strata) and if the imputation can't be done for all the variables of the questionnaire then we do it again using less detailed classes (e.g., grouping TF and TS together). We do this until all missing values are imputed.
- We also need a minimum number of units in the imputation classes, about 5 or 7 units.
- In the Caribbean context, the imputation classes can be defined as:
  - Stratum level
  - Group together TF and TS inside an IndustryGroup
  - IndustryGroup (meaning we group together TF, TS and MT)
  - Group together IndustryGroup

Reweighting approach

- We adjust the design weight of the respondents to take into account the total nonresponse units.
- The reweighting approach is based on the assumption that the non-respondents are like the respondents for the characteristics measured in the survey.
- If the reweighting is done inside the stratum, for our case, the nonresponse weight adjustment factors for a specific stratum can be defined as the ratio of the population size of the stratum to the multiplication of the number of respondents in the stratum by the initial weight of the stratum.
- Finally, the final weight for a specific stratum result of the multiplication between the initial weight of the stratum by the nonresponse weight adjustment of that stratum
- In terms of formula, it comes back to
  - $nonresponse\ weight\ adjustment = \frac{N_h}{n_{hrespondent} * Initial\ weight}$
  - $Final\ weight = initial\ weight * nonresponse\ weight\ adjustment$
- Also, an estimator of the domain total with H strata that cut across the D domains

$$\widehat{Total}_d = \sum_h \frac{N_h}{n_{hrespondent}} \sum_{s_h} \breve{y}_{krespondent}$$

where

- $\frac{N_h}{n_{hrespondent}}$ *is the final weight in the stratum h*
- $\breve{y}_{krespondent} = y_{krespondent}$ *if k is in the domain* U_d *and 0 otherwise*

And an estimator of the domain variance with H strata that cut across the D domains

$$\widehat{Variance}_d = \sum_h N_h^2 \frac{1 - {n_{hrespondent}}/{N_h}}{n_{hrespondent}} S^2_{\breve{y}s_{hrespondent}}$$

where

- $S^2_{\breve{y}s_{hrespondent}} = \frac{1}{(n_{hrespondent}-1)} \sum_{s_{hrespondent}} (\breve{y}_{krespondent} - \bar{\breve{y}}_{s_{hrespondent}})^2$

- $\bar{\breve{y}}_{s_{hrespondent}} = \sum_{s_{hrespondent}} \breve{y}_{krespondent}/n_h$
- $\breve{y}_{krespondent} = y_{krespondent}$ *if k is in the domain* U_d *and 0 otherwise*

- If the reweighting isn't done at the stratum level or if we need to combine some strata together because we have a stratum where all of the units are total nonresponse, for example, all units in Livestock_TF are total non-response and need to merge with Livestock_TS.
  - The calculation of the total need to be done at the new stratum grouping. See the example.
  - However, the calculation of the exact estimator of the domain variance is very complicated. As a proxy, we can calculate the estimator of the domain variance on the new stratum.

Example in Excel

- Example with the imputation approach

PRASC_Example_Fra
me, Sample Design ar

- Example with the reweighting approach

PRASC_Example_Fra
me, Sample Design ar