Sampling and estimation procedures in business surveys: a discussion of some specific features

David Haziza

Université de Montréal
&
CREST-ENSAI

Statistics: For Businesses, About Businesses
Royal Statistical Society
London, England

February 19, 2013

What makes business statistics special?

- Rivière (2002): excellent overview of different aspects of statistical processing and analysis of special features of business statistics
  - Heterogeneity of the universe
  - Sampling frame
  - Data collection
  - Classification of units
  - Sample coordination
  - Statistical processing: editing and imputation

OUTLINE

- Sample coordination
- Nonresponse/imputation
  - Construction of weighting classes
  - Composite imputation
- Treatment of influential units
  - What is an influential unit?
  - How measure the influence of a unit?
  - Type II winsorization

Sample coordination

- Sample coordination is required:
  - Maximize the overlap between two consecutive waves of the same survey to have accurate estimates of evolutions and reduce costs → positive coordination
  - Minimize the overlap when many surveys sample from the same Business Register in order to reduce response burden → negative coordination
- Consecutive waves of the same survey: coordination would be fairly simple if stratification is the same and the population is unchanged.
- In practice, important issues arise:
  - stratification may vary between surveys;
  - some units move from a stratum to another;
  - births;
  - deaths;
  - splits;
  - mergers;
- Samples must be updated to reflect those changes.

Sample coordination

- Several techniques used in practice are based on the concept of Permanent Random Numbers (PRN)
- Brewer, Early and Joyce (1972): use of PRN in connection with Poisson sampling
- PRN can be used for positive and negative coordination
- Each unit is assigned a random number X_i drawn from a uniform distribution, $U(0, 1)$.
- PRN can be used in connection with simple random sampling without replacement, Poisson sampling or fixed size proportional-to-size sampling; Ohlsson (1995, 1998, 1999).
- For coordination over time
  - persistants keep their PRN
  - PRN are generated for births
  - PRN are withdrawn from the list for deaths

Sample coordination

- Recent work for sample coordination in connection with Poisson sampling; Nedyalkova, Qualité and Tillé (2009a, 2009b) and Qualité (2012)
- Used in Switzerland
- Poisson sampling make things easier!
- We can easily deal with births, deaths, mergers, splits, etc.
- Greatly simplifies variance estimation

Sample coordination

- Statistics Canada is currently undertaking a major integration project for its Business Statistics surveys: Integrated Business Statistics Program (IBSP); see Turmelle, Godbout and Bosa (2012).
- Goal of IBSP: provide a common survey framework for the various business surveys conducted at Statistics Canada
- By 2016, over 100 surveys will be integrated in this new harmonized framework.
- Build around two principles:
  - Increased standardization of methods, systems, questionnaires developed around the extensive use of administrative data
  - Greater automation to achieve similar or better quality at lower processing costs

Sample coordination

- The new sampling design: two-phase stratified sampling design
- First-phase:
- the population of businesses is divided into H strata (Geography x Industry x Size)
- A sample will be selected in each stratum according to Bernoulli sampling with probability π_h
- Bernoulli design: random sample size
- Goals of the first-phase: select a large sample covering all industries
  - to verify and update activity status, industrial classification and contact data
  - collect additional information such as commodities; enable to efficiently produce good quality estimates for variables that are not available on the BR.
- For two consecutive first-phase waves: positive coordination

Sample coordination

- Second-phase:
- a subsample is selected from the first-phase sample according to stratified (Geography x Industry x Size) Bernoulli sampling
- Full questionnaire is administered to the selected units
- For two consecutive second-phase waves: negative coordination to control the response burden
- This new design will make sample coordination and variance estimation much simpler

Weighting classes

- Treatment for unit nonresponse: weight adjustment procedure within weighing classes
- The sample is partitioned into C mutually disjoint weighting classes
- Weight adjusted for nonresponse:

$$w_i = d_i/\hat{p}_c, \text{ for } i \text{ in class } c$$

  where $\hat{p}_c$ is the (weighted or unweighted) response rate in class c.
- Underlying nonresponse model: within a class, all the units have equal response probabilities.
- Weighting system adjusted for nonresponse: $\{w_i; i \in s_r\}$.
- Adjusted estimator (Propensity Score Adjusted estimator) of $Y = \sum_{i \in U} y_i$:

$$\hat{Y}_{PSA} = \sum_{i \in s_r} w_i y_i.$$

How Should We Form the Classes?

- p_i : true response probability for business i
- Asymptotic nonresponse bias of $\hat{Y}_{PSA}$:

$$Bias\left(\hat{Y}_{PSA} \mid s\right) \approx \sum_{c=1}^{C} \overline{p}_c^{-1} \sum_{i \in U_c} \left(p_i - \overline{p}_c\right)\left(y_i - \overline{y}_c\right),$$

  where $\overline{p}_c = N_c^{-1} \sum_{i \in U_c} p_i$ and $\overline{y}_c = N_c^{-1} \sum_{i \in U_c} y_i$.
- Bias of $\hat{Y}_{PSA}$ is small if $\overline{p}_c$ is large in each class, as expected.
- Bias of $\hat{Y}_{PSA}$ vanishes when the covariance between the response probability and the variable of interest is equal to 0 within each class.
- In practice, we form classes that are homogeneous with respect to response probabilities and/or to the values of the y-variable.

How Should We Form the Classes?

- Forming homogeneous classes with respect to response probabilities requires modeling the response probability, whereas forming homogeneous classes with respect to the y-variable requires postulating a model for the y-variable (i.e., assuming a prediction model).
- Survey statistician prefer modeling the response probabilities because most survey collect multiple characteristics. Modeling the y-variables requires assuming a different model for each survey variable.
- Modeling the response probabilities, by contrast, requires a single model.

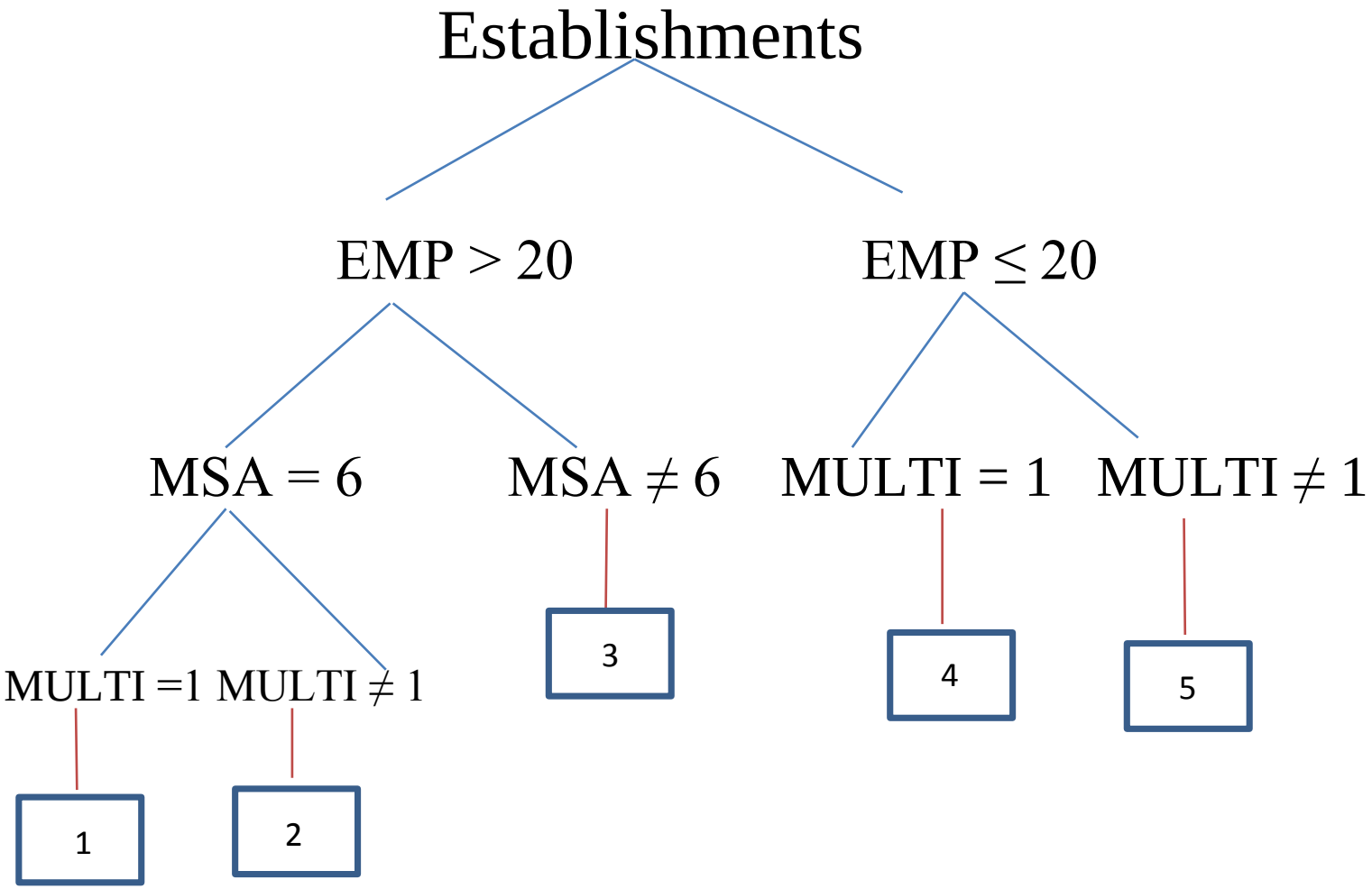
The score method for forming weighting classes

- Studied by Little (1986), Eltinge and Yansaneh (1997) and Haziza and Beaumont (2007).
- The steps for forming the classes are as follows:
  - **Step 1**: Model the response probability using a logistic regression model, which leads to the creation of the score $\hat{p}_i$ for unit i.
  - **Step 2**: Form the classes based on the estimated response probabilities, $\hat{p}_i$; e.g., equal quantile method.
  - **Step 3**: Perform weight adjustment within each class.
- Nonparametric in nature $\rightarrow$ offers some robustness if the form of logistic model is misspecified.

Using regression trees

- Issues with the score method; see Phipps and Toth (2012): Identification of interpretable weighting cells
  - Establishments in the same weighting cell may have very different characteristics
  - Logistic model in the score method: may fail to include predictors that fully accounts for curvature and interactions $\rightarrow$ lack of fit.
- Phipps and Toth (2012): use regression trees to obtain mutually disjoint cells.
- They applied the method to data from the Occupational Employment Statistics Survey.

Using regression trees



Composite imputation

- Composite imputation: more than a single imputation method is used to impute missing values for a variable of interest.
- Choice of imputation methods: depends on the availability of auxiliary variables
- Example:
  - If historical information is available, use historical imputation;
  - If another auxiliary variable is available, use ratio imputation;
  - If no auxiliary imputation is available, use mean imputation.
- Variance estimation in the presence of imputed data: widely studied in the literature; e.g., Lee, Rancourt and Särndal (2002) and Haziza (2009).
- What about variance estimation in the presence of composite imputation?

Composite imputation

- Imputed estimator of a total, $Y = \sum_{i \in U} y_i$:

$$\hat{Y}_I = \sum_{i \in s_r} d_i y_i + \sum_{i \in s_m} d_i y_i^*,$$

  where y_i^* denotes the imputed value for missing y_i.
- Linear imputation method: the imputed value y_i^* can be written as

$$y_i^* = a_{0i} + \sum_{k \in s_r} a_{ki} y_k$$

  for some coefficients a_{0i} and a_{ki}.
- Satisfied by most common imputation methods:
  - mean imputation; ratio imputation; regression imputation; historical imputation; random hot-deck imputation and nearest-neighbour imputation
- Consequence: $\hat{Y}_I = W_0 + \sum_{i \in s_r} (d_i + W_i) y_i$

Composite imputation

- Useful in the context of variance estimation;
- Särndal (1992) method for variance estimation;
- Assumes an imputation model. For example,

$$E(y_i) = \mathbf{z}_i'\boldsymbol{\beta}; \qquad V_m(y_i) = \sigma^2 c_i.$$

- Total variance of $\hat{Y}_I$ can be expressed as

$$V_{tot} = V_{SAM} + V_{NR} + V_{MIX}.$$

- The terms V_{NR} and V_{MIX} are computed under the imputation model.
- For linear imputation methods, estimators of V_{NR} and V_{MIX} can be obtained in a straightforward fashion; see Beaumont and Bissonnette (2011).
- Methodology behind SEVANI developed at Statistics Canada.

What is an influential unit?

- It is a legitimate unit which is part of the finite population:
  - The distribution of some variables of interest is highly skewed (heavy right tail)
  - An influential unit may represent other similar units in the nonsampled part.
- Important problem in business surveys
- An influential unit is not an error
  - gross error, unity measure error
  - Errors are detected at the editing stage and are treated either manually or by some for of imputation.
- We assume that errors were all detected and treated at the editing stage.
- Problem of influential units: can be reduced at the design stage (stratified sampling with take-all strata) but cannot be completely eliminated (e.g. stratum jumpers)

What is an influential unit?

Definition

A unit is influential if it has a significant impact on the sampling error, $\hat{\theta} - \theta$.

- Classical estimators (expansion and calibration estimators) are highly sensitive to influential units.
- In other words, including or excluding an influential unit from the calculations can have a dramatic impact on their magnitude $\rightarrow$ very large variance.
- It's not a problem of bias since, in the absence of nonsampling errors, classical estimators are (asymptotically) unbiased.
- Objective: reduce the influence of units that have a large influence $\rightarrow$ biased but more stable estimators .
- Treatment of influential units: compromise between bias and variance.
- One option: Winsorization

How measure the influence?

- U: finite population of size N.
- Parameter to be estimated: $Y = \sum_{i \in U} y_i$
- s: sample of size n selected using a sampling design $p(s)$.
- $d_i = \pi_i^{-1}$: sampling (design) weight of unit i.
- I_i: sample selection indicator such that $I_i = 1$ if $i \in s$ and $I_i = 0$, otherwise.
- Expansion estimator of Y: $\hat{Y}_\pi = \sum_{i \in s} d_i y_i$.
- Influence (or impact) of sample unit i on the expansion estimator:

$$B_i^\pi (I_i = 1) = E_p(\hat{Y}_\pi | I_i = 1) - Y$$

- $B_i^\pi (I_i = 1)$: conditional bias of the expansion estimator associated with unit i
- see Moreno-Rebollo, Muñoz-Reyes and Muñoz-Pichardo (1999) and Beaumont, Haziza and Ruiz Gazen (2013).

How measure the influence?

- For any sampling design, we have:

$$\begin{aligned} B_i^{\pi}(I_i = 1) &= E_p(\hat{Y}_{\pi} - Y | I_i = 1) \\ &= \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j. \end{aligned}$$

- The conditional bias generally depends on the second-order inclusion probabilities, π_{ij}.
- The conditional bias is, in general, unknown $\rightarrow$ it must be estimated
- If $\pi_i = 1$ then

$$B_i^{\pi}(I_i = 1) = 0.$$

How measure the influence?

- The population U is partitioned into H mutually exclusive strata $U_1, \ldots, U_H$ of size $N_1, \ldots, N_H$.
- Stratum totals: $Y_1, \ldots, Y_H$
- Relationship: $Y = \sum_{h=1}^{H} Y_h$
- A sample s_h, of size n_h, is selected from U_h according to simple random sampling without replacement.
- Relationship:

$$\sum_{h=1}^{H} \hat{Y}_{h,\pi} = \hat{Y}_{\pi},$$

where $\hat{Y}_{h,\pi} = \sum_{i \in s_h} d_i y_i$ and $d_i = N_h / n_h$.

How measure the influence?

- Stratified simple random sampling without replacement:

$$B_i^{\pi}(I_i = 1) = \frac{N_h}{(N_h - 1)}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{Y}_h) \text{ for } i \in U_h,$$

where $\bar{Y}_h = Y_h/N_h$ is the population mean in stratum h.

- Estimator of the conditional bias:

$$\hat{B}_i^{\pi}(I_i = 1) = \frac{N_h}{(N_h - 1)}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_h) \text{ for } i \in U_h,$$

where $\bar{y}_h = \sum_{i \in s_h} y_i/n_h$ is the sample mean in stratum h.

- Bernoulli sampling within stratum:

$$B_i^{\pi}(I_i = 1) = (\pi_h^{-1} - 1)y_i = (\pi_h^{-1} - 1)(y_i - 0) \text{ for } i \in U_h$$

Type II winsorization in stratified sampling

- Apply winsorization independently within each stratum
- Type II winsorization : the variable y after winsorization is given by

$$\tilde{y}_i = \begin{cases} y_i \text{ si } y_i \leq K_h \\ f_h y_i + (1 - f_h) K_h \text{ si } y_i > K_h \end{cases}$$

  if $i \in U_h$, where $f_h = n_h / N_h$.
- Special case: $f_h = 0 \Rightarrow$ Type I winsorization
- Type II winsorized estimator in stratum h:

$$\hat{Y}_{h,win} = \sum_{i \in s_h} d_i \tilde{y}_i = \sum_{i \in s_h} \tilde{d}_i y_i,$$

  where

$$\tilde{d}_i = 1 + \left(\frac{N_n}{n_h} - 1 \right) \frac{\min(y_i, K_h)}{y_i}.$$

Type II winsorization in stratified sampling

- Issue: determine the cut-off points $K_1, \ldots, K_H$.
- One option: determine the cut-off that minimize the estimated mean square error of the winsorized estimator; see Kokic and Bell (1994) and Rivest and Hurtubise (1995).
- Generally complex to implement.
- Alternative option: Beaumont, Haziza and Ruiz-Gazen (2013)
  - In stratum h, determine K_h that minimizes

$$\max\left\{|\hat{B}_i^{win}(I_i = I)|\,;\; i \in s_h\right\},$$

where $\hat{B}_i^{win}(I_i = I) = E_p(\hat{Y}_{h,win}|I_i = 1) - Y_h$ denotes the conditional bias of the winsorized estimator for unit i.

- Under this criteria, the winsorized estimator in stratum h is given by

$$\hat{Y}_{h,win}(K_{h,opt}) = \hat{Y}_{h,\pi} - \frac{1}{2}\left(\hat{B}_{min}^{\pi} + \hat{B}_{max}^{\pi}\right).$$

Type II winsorization in stratified sampling

- Robust estimator of the overall total Y:

$$\hat{Y}_{agr} = \sum_{h=1}^{H} \hat{Y}_{h,win}(K_{h,opt}).$$

- Consistency is satisfied.
- Issue: $\hat{Y}_{agr}$ is the sum of (negative) biased estimators $\rightarrow$ may be heavily biased; see Rivest and Hidiroglou (2004)
- One possible option:
  - Apply Type II winsorization independently within each stratum and obtain the winsorized estimators, $\hat{Y}_{1,win}(K_{1,opt}), \ldots, \hat{Y}_{H,win}(K_{H,opt})$.
  - Independently, apply Type II winsorization and obtain a winsorized estimator of the overall total Y: $\hat{Y}_{win}(K_{opt})$
  - However, the consistency relationship is no longer satisfied as

$$\sum_{h=1}^{H} \hat{Y}_{h,win}(K_{h,opt}) \neq \hat{Y}_{win}(K_{opt}).$$

Type II winsorization in stratified sampling

- Solution: force consistency
- Find final winsorized estimates $\hat{Y}^*_{1,win}(K_{1,opt}), \ldots, \hat{Y}^*_{H,win}(K_{H,opt})$ as close as possible to the initial winsorized estimates $\hat{Y}_{1,win}(K_{1,opt}), \ldots, \hat{Y}_{H,win}(K_{H,opt})$, so that the calibration

$$\sum_{h=1}^{H} \hat{Y}^*_{h,win}(K_{h,opt}) = \hat{Y}_{win}(K_{opt}),$$

  is satisfied.
- Similar to calibration for weighting

Type II winsorization in stratified sampling

- More specifically, find the final winsorized estimates $\hat{Y}^*_{1,win}(K_{1,opt}), \ldots, \hat{Y}^*_{H,win}(K_{H,opt})$ so that

$$\sum_{h=1}^{H} \frac{1}{q_h} \frac{\left\{ \hat{Y}^*_{h,win}(K_{h,opt}) - \hat{Y}_{h,win}(K_{h,opt}) \right\}^2}{\hat{Y}_{h,win}(K_{h,opt})}$$

  is minimized such that

$$\sum_{h=1}^{H} \hat{Y}^*_{h,win}(K_{h,opt}) = \hat{Y}_{win}(K_{opt})$$

  is satisfied; see Favre-Martinoz, Haziza and Beaumont (2013).

- q_h : coefficient assigned to stratum h
- Assign a small value of q_h if we do not wish to modify the initial winsorized estimate $\hat{Y}_{h,win}(K_{h,opt})$ to a large extent.

A simulation study

- We generated a population of size $N = 5000$ consisting of 5 strata
- In each stratum, we generated a variable of interest y from a log-normal distribution with parameter log(2000) and 1.5.
- From the population, we selected 5000 samples according to stratified simple random sampling without replacement.

Stratum	1	2	3	4	5
N_h	2000	1500	1000	400	100
f_h	0.01	0.05	0.1	0.2	0.8

- We computed the Monte Carlo percent relative bias:

$$RB(\hat{\theta}) = 100 \times E_{MC}(\hat{\theta}_{win} - \theta)/\theta$$

- We computed the Relative Efficiency:

$$RE(\hat{\theta}_{win}) = 100 \times MSE_{MC}(\hat{\theta}_{win})/MSE_{MC}(\hat{\theta}_{\pi}).$$

A simulation study

		$\hat{Y}_{agr}$	$\hat{Y}_{win}(K_{opt})$	$\hat{Y}_{win}(K_{opt})$
Overall estimator		$-8.9(87)$	$-4.4(77)$	$-4.4(77)$
		$\hat{Y}_{h,win}(K_{h,opt})$	$\hat{Y}^{*}_{h,win}(K_{h,opt})$	
			$\mathbf{q} = (1,1,1,1,1)$	$\mathbf{q} = (1,1,1,1,0.1)$
Stratum	1	$-16.9(65)$	$-12.7(67)$	$-12.0(69)$
	2	$-10.3(80)$	$-5.9(80)$	$-4.7\ (81)$
	3	$-9.2(54)$	$-7.4(55)$	$-3.9(56)$
	4	$-8.2(74)$	$-3.7(77)$	$-3.3(79)$
	5	$-1.4(97)$	$3.3(120)$	$-0.72(97)$

Other topics

- Selective editing: who should we edit?
  - Score function approach: e.g., Lawrence and McKenzie (2000), Hedlin (2003) and Brion (2009)
- Chasing the nonrespondents: who should we chase?
  - Berger (2009): Monthly Inquiry into the Distribution and Services Sector conducted at Office for National Statistics;
  - Brodie (2012).
- Small area estimation: presence of influential units
  - e.g., Sinha and Rao (2009), Dongmo Jiongo, Haziza and Duchesne (2012), Chambers, Chandra, Salvati and Tzavidis (2013), Tzavidis (2012, Dutch Structural Business Survey).