Investigation of Variance Estimators for the Survey of Business Owners (SBO)

Marilyn Balogh and Sandy Peterson
U.S. Census Bureau

November 5, 2013

Outline

- Background on SBO
- Variance Estimation Methodology
  - Random group (simple and stratum-specific)
  - Delete-a-group Jackknife (simple and stratum-specific)
  - Stratified Jackknife
- Simulation Study
- Results
- Conclusion

Background on SBO

- Part of the Economic Census taken every 5 years for years ending in “2” and “7”
- The only comprehensive, regularly collected data for businesses and business owners by
  - Gender
  - Race
  - Ethnicity (Hispanic origin of any race)
  - Veteran status

Background on SBO

- SBO universe:
  - 9 sampling frames based off modeled likelihoods
  - stratify by frame, state, industry code, and employment status (68,585)
- Firms are selected with certainty or are subjected to systematic sampling

Background on SBO

- Hot-deck donor imputation for unit and item non-response
- Calculate estimates using Horvitz-Thompson estimator
- Estimates sampling error using the random group (RG) variance estimator
  - 10 non-certainty random groups
  - fpc adjustment factor

Variance Estimation Methodology

- Three variance estimators:
  - Random group (RG)
    - simple and stratum-specific
  - Delete-a-group jackknife (DAG)
    - simple and stratum-specific
  - Stratified jackknife (SJK)

Random Group and Delete-a-Group Jackknife Methods

- Divides the non-certainty firms into *R* random groups
- Creates *R* replicate estimates
- Calculates the simple variance
- 2 reweighting procedures (simple and stratum-specific)

RG simple method

- The replicate-*r* weight is:

$$w_{ri} = \begin{cases} R * w_i' & i \,\epsilon \text{ replicate group } r \\ 0 & \text{otherwise} \end{cases}$$

  - where:
    - w_i' is the fpc-adjusted sampling weight of the unit
    - R is the number of non-certainty random groups

RG stratum-specific method

- The replicate-*r* weight is:

$$w_{rhi} = \begin{cases} \frac{n_h}{n_{hr}} * w'_{hi} & i \in \text{replicate group } r \\ 0 & \text{otherwise} \end{cases}$$

  - Where:
    - w'_{hi} is the fpc-adjusted sampling weight of unit *i* in stratum *h*
    - n_h is the total number of non-certainty sampled units in stratum *h*
    - n_{hr} is the total number of non-certainty sampled units in stratum *h* and replicate group *r*

RG Variance

- The RG variance for any estimate $\hat{\theta}$ is:

$$\widehat{var}_{RG}(\hat{\theta}) = \frac{1}{R(R-1)} \sum_{1}^{R} \left(\hat{\theta}_{r(RG)} - \hat{\theta}'\right)^2$$

  - where:
    - $\hat{\theta}_{r(RG)}$ is the RG replicate *r* estimate
    - $\hat{\theta}'$ is the fpc adjusted full sample estimate

DAG Simple Method

- The replicate-*r* weight is:

$$w_{ri} = \begin{cases} 0 & i \in replicate\ group\ r \\ \frac{R}{R-1} * w_i' & otherwise \end{cases}$$

  - Where:
    - w_i' is the fpc-adjusted sampling weight
    - R is the number of random groups

DAG stratum-specific method

- Assume all stratum sample sizes are large $(n_h \geq 5)$
- For SBO, this assumption does not hold, so we used the extended DAG method[1]
  - Developed by Phil Kott in the "The Delete-a-Group Jackknife" (Journal of Official Statistics)
  - Assume all stratum sample sizes have at least 2 units

DAG stratum-specific method

- When $1 < n_h < R$, then the replicate-*r* weight is:

$$w_{rhi} = \begin{cases} w'_{hi} & S_{hr} \text{ is empty} \\ w'_{hi} * (1 - [n_h - 1]Z) & i \in S_{hr} \\ w'_{hi} * (1 + Z) & \text{otherwise} \end{cases}$$

  - where:
    - S_{hr} is the set of n_{hr} non-certainty sampled firms in stratum *h* and random group *r*
    - $Z^2 = R/[(R-1)n_h(n_h-1)]$

DAG stratum-specific method

- When $n_h \geq R$, then the replicate-*r* weight is:

$$w_{rhi} = \begin{cases} 0 & i \in S_{hr} \\ w'_{hi} * \left(\frac{n_h}{n_h - n_{hr}}\right) & \text{otherwise} \end{cases}$$

DAG Variance

- The DAG variance for any estimate $\hat{\theta}$ is:

$$\widehat{var}_{DAG}(\hat{\theta}) = \frac{R-1}{R}\sum_{1}^{R}\left(\hat{\theta}_{r(DAG)} - \hat{\theta}'\right)^2$$

  - where:
    - $\hat{\theta}_{r(DAG)}$ is the DAG *r* replicate estimate
    - $\hat{\theta}'$ is the fpc adjusted full sample estimate

Stratified Jackknife Method

- Constructs one replicate estimate per sampling unit
  - Drop one unit at a time from the stratum and multiply the remaining units in the stratum by $n_h/(n_h - 1)$
  - Assume 2 non-certainty sampled units within each stratum
- Calculate the simple variance

SJK Method

- The replicate-*k* weight is:

$$w_{khi} = \begin{cases} w_i & \text{if unit } i \text{ is not in stratum } h \\ \frac{n_h}{n_h - 1} w_i & \text{if unit } i \text{ is in stratum h but not unit } k \\ 0 & \text{otherwise} \end{cases}$$

  - where:
    - w_i is the sampling weight of unit *i*
    - n_h is the number of sampled units in stratum *h*

SJK Variance

- The SJK variance for any estimate $\hat{\theta}$ is:

$$\widehat{var}_{SJK}(\hat{\theta}) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{k=1}^{n_h} (1 - f_{hi}) \left(\hat{\theta}_{k(SJK)} - \hat{\theta}\right)^2$$

  - where:
    - f_{hi} is equal to the probability of selection for firm *i* in stratum *h*
    - *H* is the number of strata
    - $\hat{\theta}_{r(SJK)}$ is the SJK replicate estimate
    - $\hat{\theta}$ is the full sample estimate

Simulation Study

- Created a simulated population
  - Selected 5 states
    - Florida, Georgia, Kansas, New York, and North Dakota
  - Assigned race, gender, ethnicity, and veteran status
- Selected 5,000 different stratified systematic samples

Simulation Study

- Assigned sampled units to 10 non-certainty random groups
- Calculated the 5 variance estimators:
  - RG simple (RG_S)
  - RG stratum-specific (RG_ST)
  - DAG simple (DAG_S)
  - DAG stratum-specific extended (DAG_ST)
  - Stratified Jackknife (SJK)

Simulation Study

- Calculated the relative bias (RB) and the coefficient of variation (CV)

$$RB_{METH} = \frac{\overline{\widehat{var}_{METH}(\hat{\theta})}}{var(\hat{\theta})} - 1$$

$$CV_{METH} = \frac{\sqrt{\frac{1}{5000}\sum_{i=1}^{5000}[\widehat{var}_{METH}(\hat{\theta}_i) - var(\hat{\theta})]^2}}{var(\hat{\theta})}$$

Results

- Best variance estimator should have the RB and the CV near zero
- Ran a sign test on the pairwise differences of each method's CV across all domains by estimate type
- H_0: The median of X_i equals the median of Y_i for all i
  - Paired the data for 2 methods within each domain and calculated the difference of the CV
  - Calculated the test statistic T, which is the number of times the difference was greater than zero

CV Sign Tests Results

- Median of CVs of SJK method is smaller than median of other methods
- Median of CVs of RG_ST method is smaller than median of other methods, except SJK
- Median of CVs of DAG_ST method is smaller than the simple methods

Relative Bias Results

Table 2: Relative biases for the firm count by demographic characteristic and variance estimator for all firms within New York

Demographic Characteristic	Relative Bias				
	RG_S	RG_ST	DAG_S	DAG_ST	SJK
All firms	0.174	31.417	0.174	-0.997	-0.988
Female	-0.032	0.048	-0.032	-0.040	-0.012
Male	-0.069	0.222	-0.069	-0.086	-0.076
Hispanic	-0.050	-0.049	-0.050	-0.058	-0.047
Non-Hispanic	-0.054	1.664	-0.054	-0.136	-0.146
White	-0.245	0.414	-0.245	-0.281	-0.276
Black or African American	-0.095	-0.074	-0.095	-0.107	-0.098
AIAN	-0.078	-0.064	-0.078	-0.079	-0.049
Asian	-0.552	-0.544	-0.552	-0.555	-0.560
NHOPI	-0.011	-0.018	-0.011	-0.011	-0.020

Coefficient of Variation Results

Table 3: CVs for the firm count by demographic characteristic and variance estimator for all firms within New York

Demographic Characteristic	Coefficient of Variations				
	RG_S	RG_ST	DAG_S	DAG_ST	SJK
All firms	0.643	31.419	0.643	0.997	0.988
Female	0.449	0.445	0.449	0.444	0.014
Male	0.446	0.487	0.446	0.443	0.077
Hispanic	0.454	0.451	0.454	0.452	0.048
Non-Hispanic	0.449	1.714	0.449	0.429	0.146
White	0.429	0.535	0.429	0.439	0.276
Black or African American	0.442	0.430	0.442	0.438	0.098
AIAN	0.451	0.445	0.451	0.452	0.065
Asian	0.591	0.582	0.591	0.593	0.560
NHOPI	0.490	0.483	0.490	0.491	0.124

Real Time Results

- The SJK method generally has the lowest CVs
- Amount of time to run the SJK method is extremely high
  - For our small study sample, the SJK method took 12.6 times longer
  - For the full 2007 SBO sample, the SJK method took 73 times longer
  - SJK method would take over a month to run all the estimates

Conclusion

- SJK variance estimator was the superior method
  - Consistently produced a low CV
  - Showed little difference in RB
- Processing time for SJK method would take too long
  - Recommend future research into more efficient processing for the SJK variance estimator

Conclusion

- RG_ST showed huge fluctuations in both the RB and CV for the firm count
- No apparent difference between the RG_S, DAG_S, and DAG_ST method
- We recommend the DAG_ST variance estimator because it better handles strata with few records ($n_h \leq 5$)

Acknowledgements

- Sandy Peterson
- Maxwell Mitchell
- Jeffrey Dalzell
- Robin Gibson
- Terry Pennington
- Beth Schlein
- Meijin Ye

Contact Information

- Marilyn Balogh, Mathematical Statistician, US Census Bureau Marilyn.K.Balogh@census.gov
- Sandy Peterson, Mathematical Statistician, US Census Bureau Sandra.Peterson@census.gov
- General SBO inquiries
  Phone: 888.225.4022 or 301.763.3316
  Email: csd.sbo@census.gov