Refocusing on What We Don’t Know: A Sample Redesign to Leverage Administrative Data

Sandy Peterson, U.S. Census Bureau
Stephen Hardy, U.S. Census Bureau
Daniel Cordes, U.S. Census Bureau
Audrey Kindlon, National Center for Science and Engineering Statistics

FCSM

October 23, 2024

DRB Clearance #CBDRB-FY24-ESMD013-001

This presentation provides results of exploratory research for a survey that is sponsored in part by the National Center for Science and Engineering Statistics (NCSES) within the National Science Foundation (NSF). Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the Census Bureau, NCSES, or NSF.

The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Project No. P-7504866, Disclosure Review Board (DRB) approval number: CBDRB-FY24-ESMD013-001).

All estimates are subject to sampling error, and all comparison statements made during this presentation are significant at the 90% confidence level.

Annual Business Survey (ABS)

Annual survey of nonfarm, for-profit employer firms collecting data on:

- Business owner demographics
- Business and business owner characteristics
- Innovation activities
- Research and Development (R&D) expenditures
- Rotating content on financing, management practices, globalization, design, or climate and sustainability

ABS Sample Design

- Sample 300,000 firms each year, including 45,000 certainty firms due to large size or known R&D
- Universe stratified by expected owner demographics, state, and primary industry
- Oversample small demographic groups and high-R&D industry strata
- Select at least 5 noncertainty firms per stratum

Motivation for Sample Redesign

- Administrative data available for owner demographics
    - No longer need sample to measure demographics, so sample should be designed to measure other characteristics

- Uneven respondent burden
    - Small firms with rare combinations of characteristics have been selected in most (sometimes all) years of ABS

Simulation Research Plan

- Create 2 populations for consecutive years with fabricated "true" characteristics for all firms
- Use historical ABS responses and improved administrative data sources and methods to determine known demographics
- Evaluate 4 different sampling methods based on:
  - Stratum-level sampling rates
  - Noncertainty sample overlap from first to second year
  - Estimate bias and variance

Simulation Samples

Sample Name	Stratification	Sample Parameters	Controlled Nonselection
Original	Current	Current	No
Alternate 1 (Collapse)	Collapsed	Current	No
Alternate 2 (New Params)	Collapsed	Updated	No
Alternate 3 (Nonselection)	Collapsed	Updated	Yes

Current Stratification Cells

- Expected owner demographics (7 categories)
  - Native Hawaiian or Other Pacific Islander (NHOPI)
  - American Indian or Alaska Native (AIAN)
  - Black or African American
  - Asian
  - Hispanic
  - White female
  - White male or unclassifiable
- State (52 categories)
  - Includes DC and a category for multi-unit firms operating in multiple states
- Primary industry groups most relevant to NCSES publications (35 categories)
  - All other industries combined into one large industry group

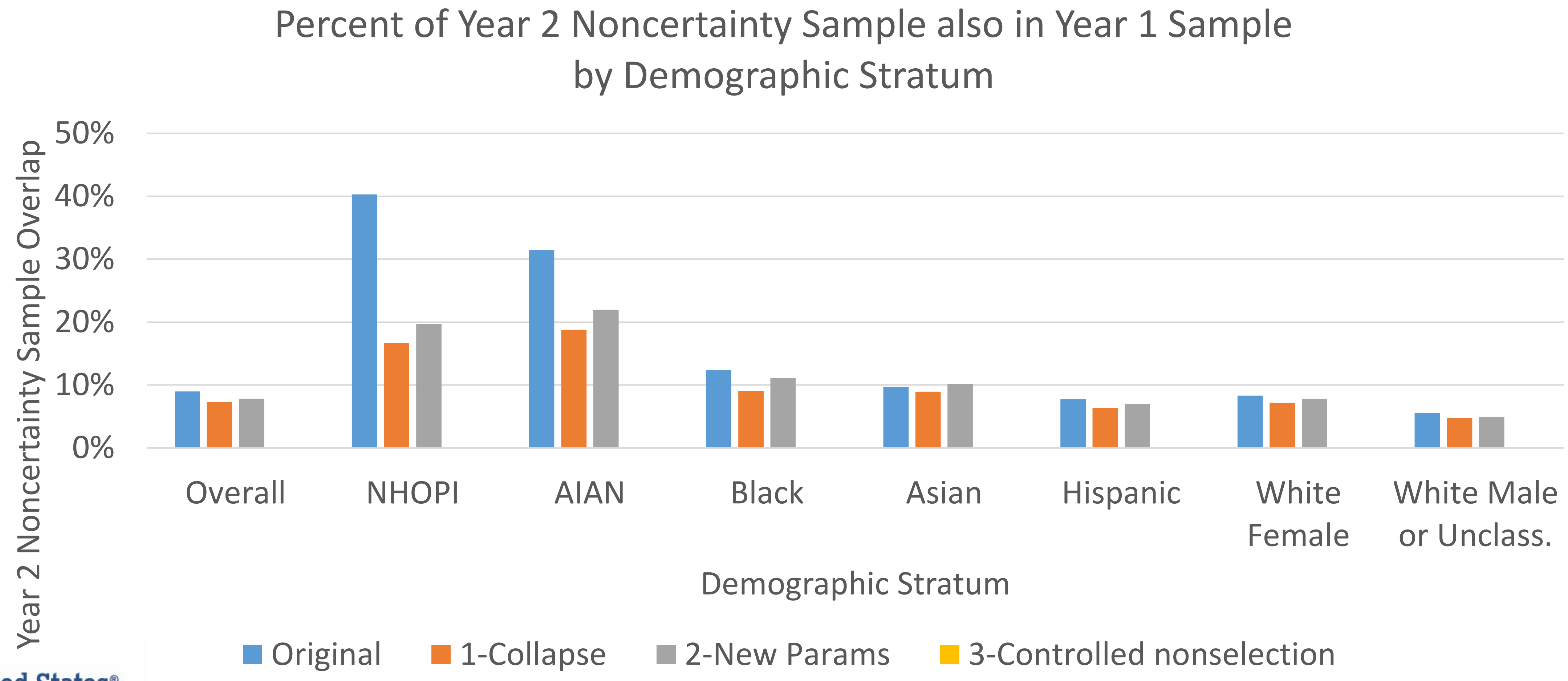
Proposed Stratification Collapsing Method

- As needed, combine strata together until all strata have at least 15 firms
  - Never combine strata from different states
  - Combine demographic groups within oversampled R&D industry
  - For all other industries, combine industries within demographic group

Stratification Collapsing Results

Stratum Size	Original Strata (and sampling rate)	Collapsed Strata (and sampling rate)
1-5 firms	3,091 (100%)	0
6-14 firms	1,520 (69%)	0
15-30 firms	1,198 (30%)	1,270 (32%)
31-50 firms	811 (26%)	1,024 (31%)
51-100 firms	829 (17%)	936 (20%)
101 or more firms	2,141 (4%)	2,162 (4%)
Total	**9,590 (4%)**	**5,392 (4%)**

Sample Overlap



DRB Clearance #CBDRB-FY24-ESMD013-001

Mean Relative Standard Error (RSE)

ABS uses extended delete-a-group jackknife variance estimation with 10 random groups and a finite-population-correction.

$$RSE\left(\hat{Y}\right) = 100 * \frac{\sigma_{\hat{Y}}}{\hat{Y}}$$

For statistical comparison between sample types, we calculate the simple mean and variance of RSE across 11 different samples:

$$var\left(RSE\left(\hat{Y}\right)\right) = \frac{1}{11}\sum\left(RSE\left(\hat{Y}\right) - \overline{RSE\left(\hat{Y}\right)}\right)^2$$

Mean Relative Standard Error (RSE)

Number of estimates where alternate sample types produced a higher Mean RSE than the original estimate.

	Total Estimates	1-Collapsed Higher RSE	2-New Parameter Higher RSE	3-Nonselection Higher RSE
0 or 1 characteristic	468	0	0	0
2 chars excl NAICS 3 & 4	1,585	22	20	22
3 chars excl NAICS 3 & 4	2,043	12	24	23
2 chars incl NAICS 3 & 4	18,946	377	433	435
3 chars incl NAICS 3 & 4	17,362	244	421	439
Overall	**40,404**	**655**	**898**	**919**

Mean Relative Standard Error (RSE)

Number of estimates where alternate sample types produced a lower Mean RSE than the original estimate.

	Total Estimates	1-Collapsed Lower RSE	2-New Parameter Lower RSE	3-Nonselection Lower RSE
0 or 1 characteristic	468	0	10	7
2 chars excl NAICS 3 & 4	1,585	1	7	9
3 chars excl NAICS 3 & 4	2,043	6	9	9
2 chars incl NAICS 3 & 4	18,946	107	206	185
3 chars incl NAICS 3 & 4	17,362	89	145	140
Overall	**40,404**	**203**	**377**	**350**

Mean Percent Bias

Percent bias for each estimate compares the true sum to the estimated sum and divides by the true sum to standardize across estimates.

$$B = \frac{\hat{Y} - Y}{Y}$$

For statistical comparison between sample types, we calculate the simple mean and variance of the Percent Bias across 11 different samples:

$$var(B) = \frac{1}{11}\sum(B - \bar{B})^2$$

Mean Percent Bias

Estimates with significant bias by sample type

	Total Estimates	Original Estimates with Bias	1- Collapse Estimates with Bias	2-New Parameter Estimates with Bias	3-Nonselection Estimates with Bias
0 or 1 characteristic	468	2	2	3	1
2 chars excl NAICS 3 & 4	1,585	21	23	25	27
3 chars excl NAICS 3 & 4	2,043	51	41	63	59
2 chars incl NAICS 3 & 4	18,946	1,162	1,165	1,037	1,033
3 chars incl NAICS 3 & 4	17,362	1,136	1,161	1,107	1,143
Overall	**40,404**	**2,372**	**2,392**	**2,235**	**2,263**

Conclusions

- New administrative data sources provide more accurate and comprehensive demographic ownership data than before
- Collapsing strata and new sampling parameters reduce respondent burden for small firms with rare characteristics
- Preliminary analysis shows minimal impacts to estimate quality using collapsed strata and new sampling parameters

Questions welcome at sandra.peterson@census.gov