Variance Estimation for Product Value Estimates in the 2017 Economic Census Under the Assumption of Complete Response

Matthew Thompson
Katherine J. Thompson
Robin Kurec

U.S. Census Bureau*

Prepared for JSM July 31-August 4, 2016

**The views expressed in this presentation are those of the authors and not necessarily those of the U.S. Census Bureau*

Research Challenge

- Produce variance estimates for "product sales" (products) collected in the 2017 Economic Census

- Preliminary research (phase 1):
  - Imputation variance (previous presentation)
  - Sampling variance
  - Post-stratification

Economic Census Background

- Not strictly a census
  - Multi-units and large single-units selected with certainty
  - Small single-units sampled

- Sampling varied by trade for the 2012 EC
  - Wholesale: Census
  - Manufacturing, Mining: Cutoff sample
  - Construction: PPS
  - All others: Stratified systematic sampling

Economic Census Background

- Not strictly a census
  - Multi-units and large single-units selected with certainty
  - Small single-units sampled

- Sampling varied by trade for the 2012 EC
  - ~~Wholesale: Census~~ **(excluded)**
  - ~~Manufacturing, Mining: Cutoff sample~~ **(excluded)**
  - Construction: PPS
  - All others: Stratified systematic sampling

Economic Census Background

- Final product estimates are produced by calibration to stratum-level receipt totals
- Samples designed for estimation, not specifically for direct variance estimation
- Not uncommon for strata to contain only 1 or 2 sampled establishments

Collapsing Strata

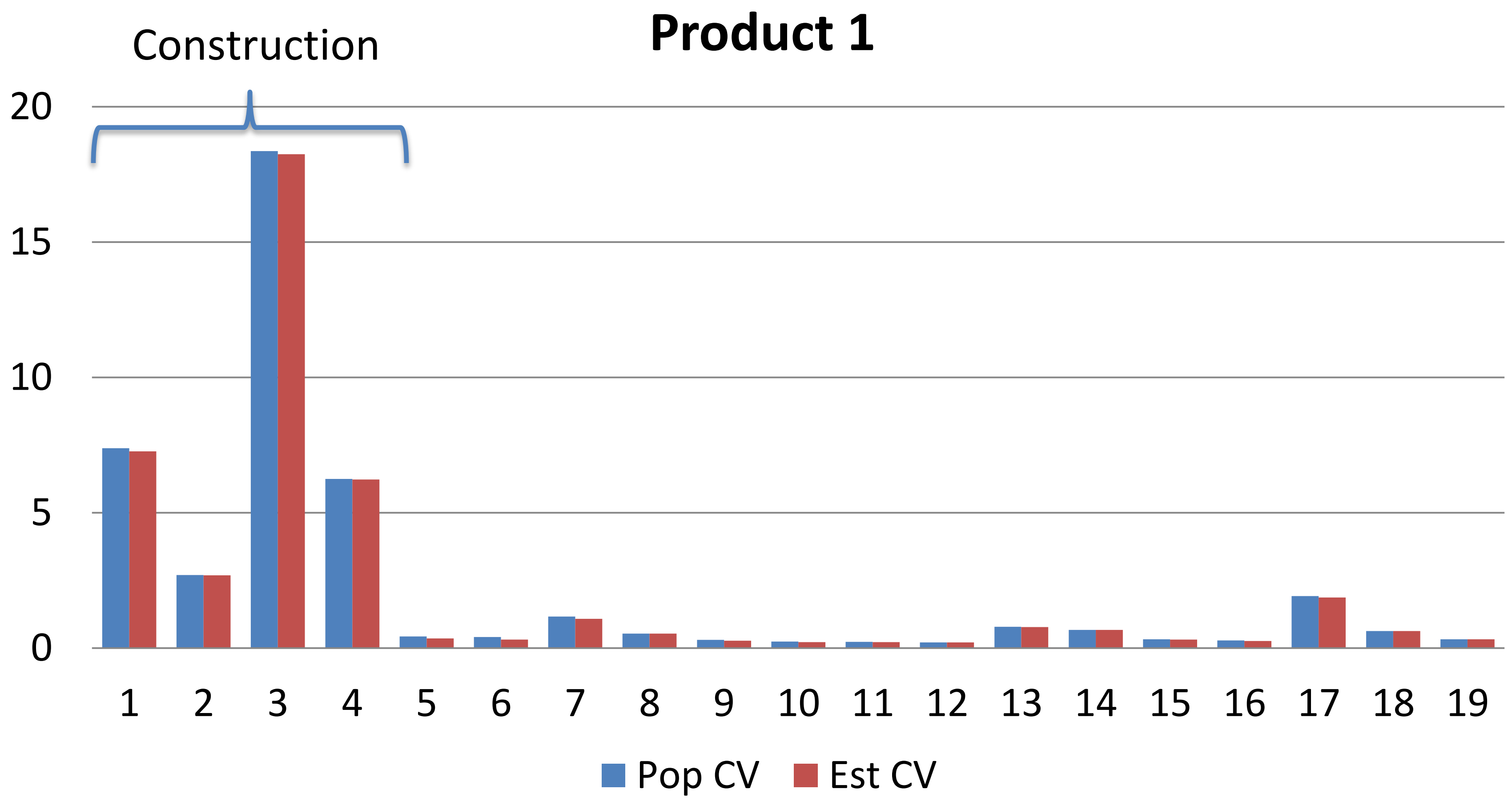
- States with similar average receipts values within an industry were combined to create strata

- Strata were chosen such that each contained at least 10 establishments

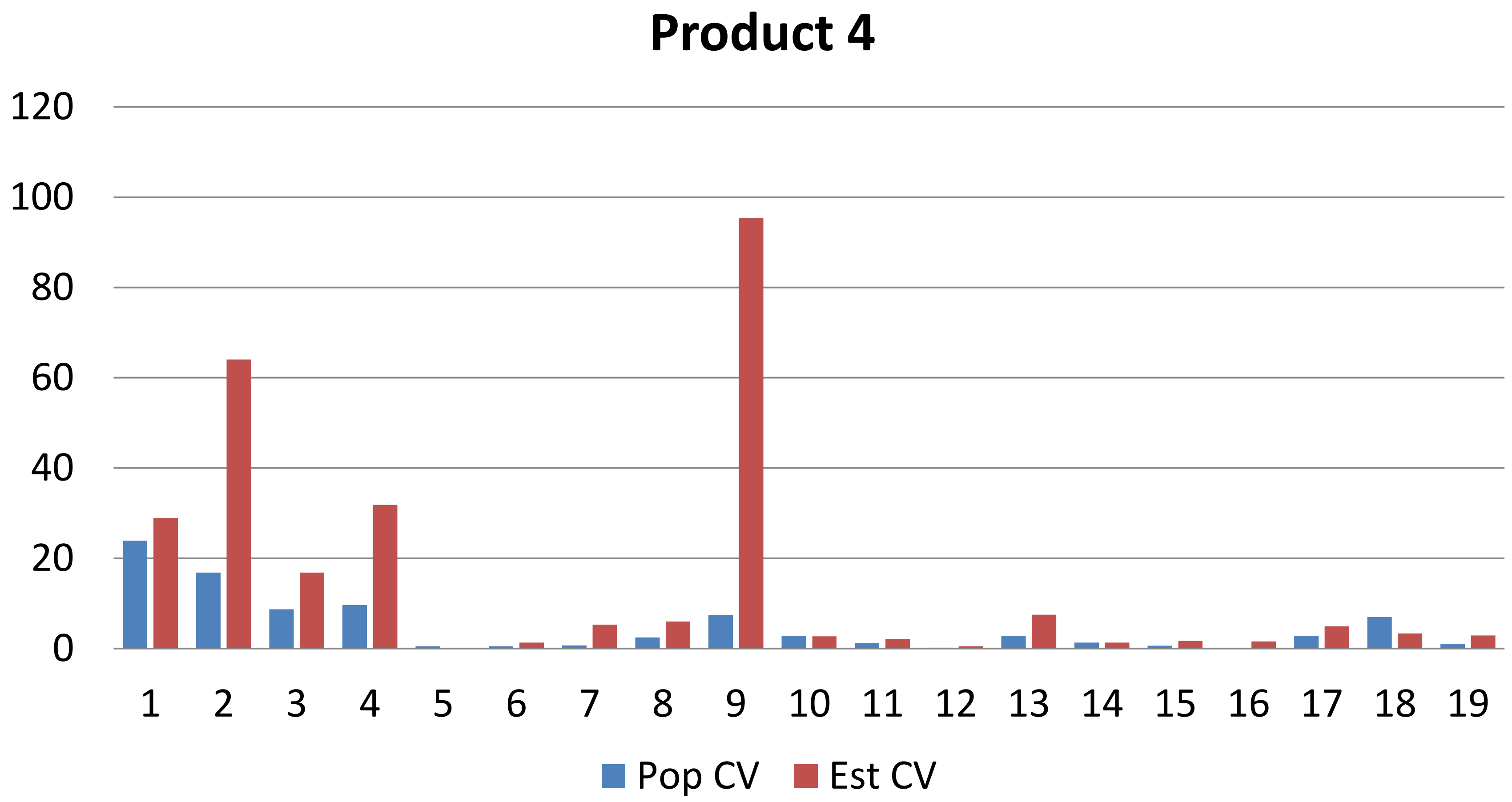
Research Data – "Populations"

- Historic Economic Census data (2012)
- Impute missing product values
- Retain "five" products per industry
- "Expand" sample to population
- Draw 5,000 Stratified WOR-SRS samples

Population CV vs Estimated CV



Population CV vs Estimated CV

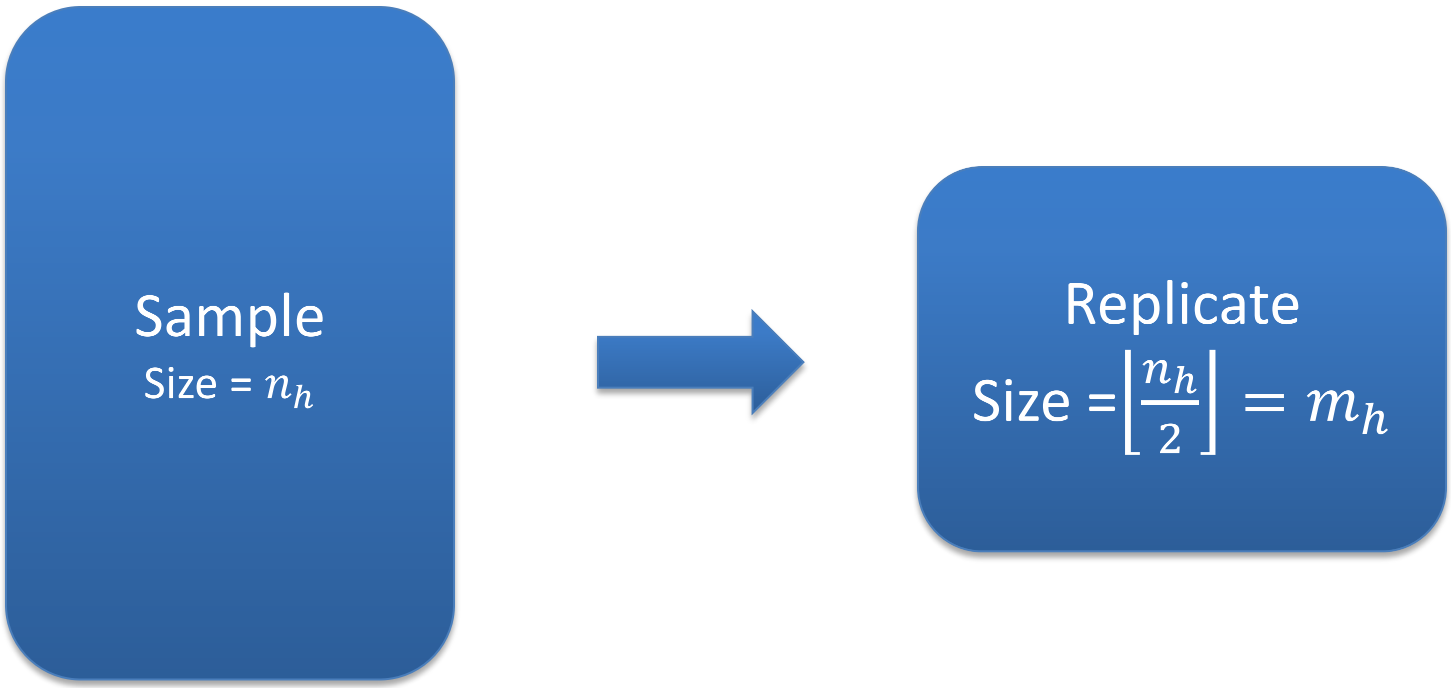


Variance Estimation Methods

- 3 design-based replication methods
  - Chipperfield-Preston
  - Mirror Match
  - Without Replacement Bootstrap

- Finite Population Bayesian Bootstrap

Chipperfield-Preston (CHIP)



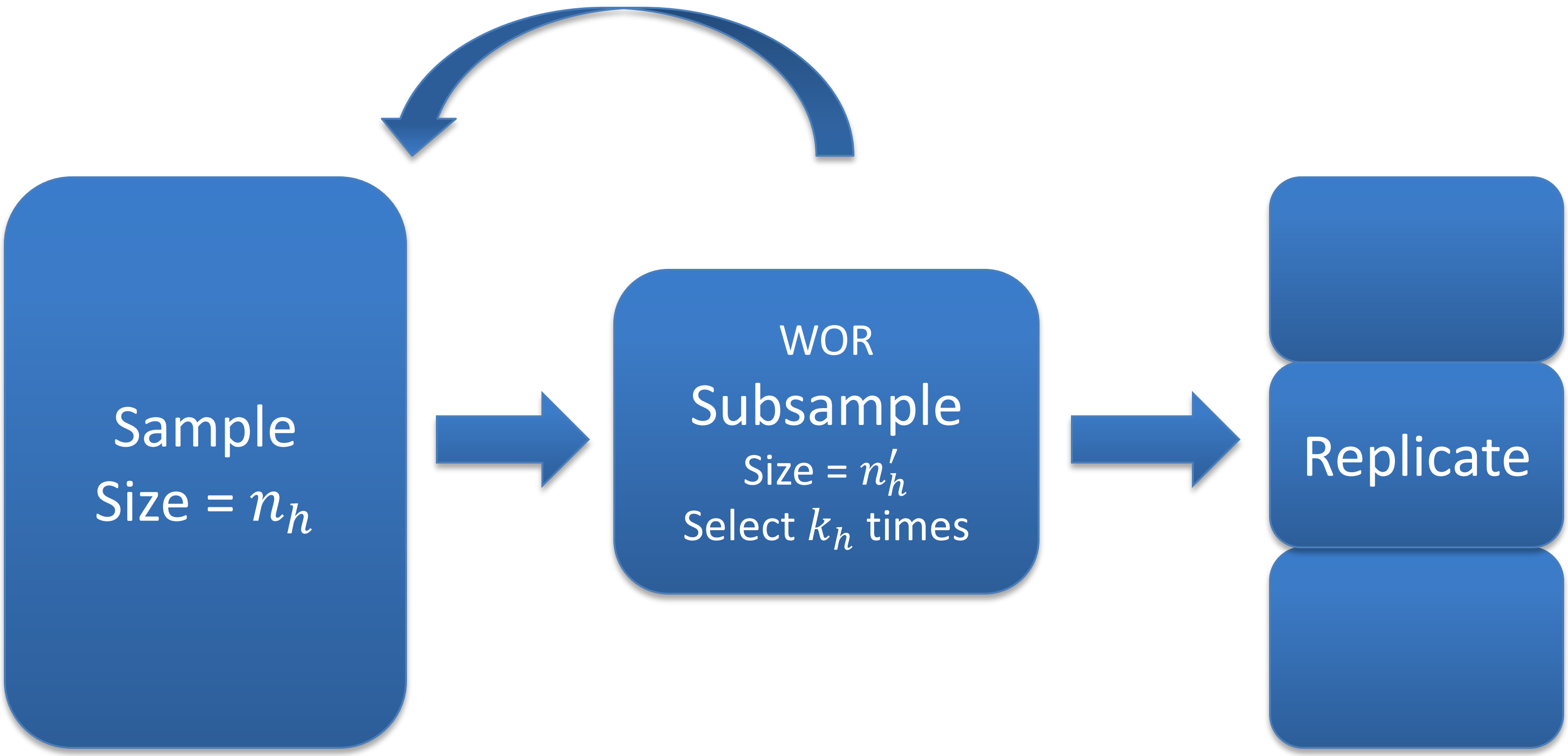
Chipperfield-Preston cont'd

- Adjust the weights for units in replicate

$$w_{hi}^{*} = w_{hi}\left(1 - \gamma_h + \gamma_h\left(\frac{n_h}{m_h}\right)\right)$$

$$\gamma_h = \sqrt{(1 - f_h)m_h/(n_h - m_h)}$$

Mirror Match (MM)



Mirror Match cont'd

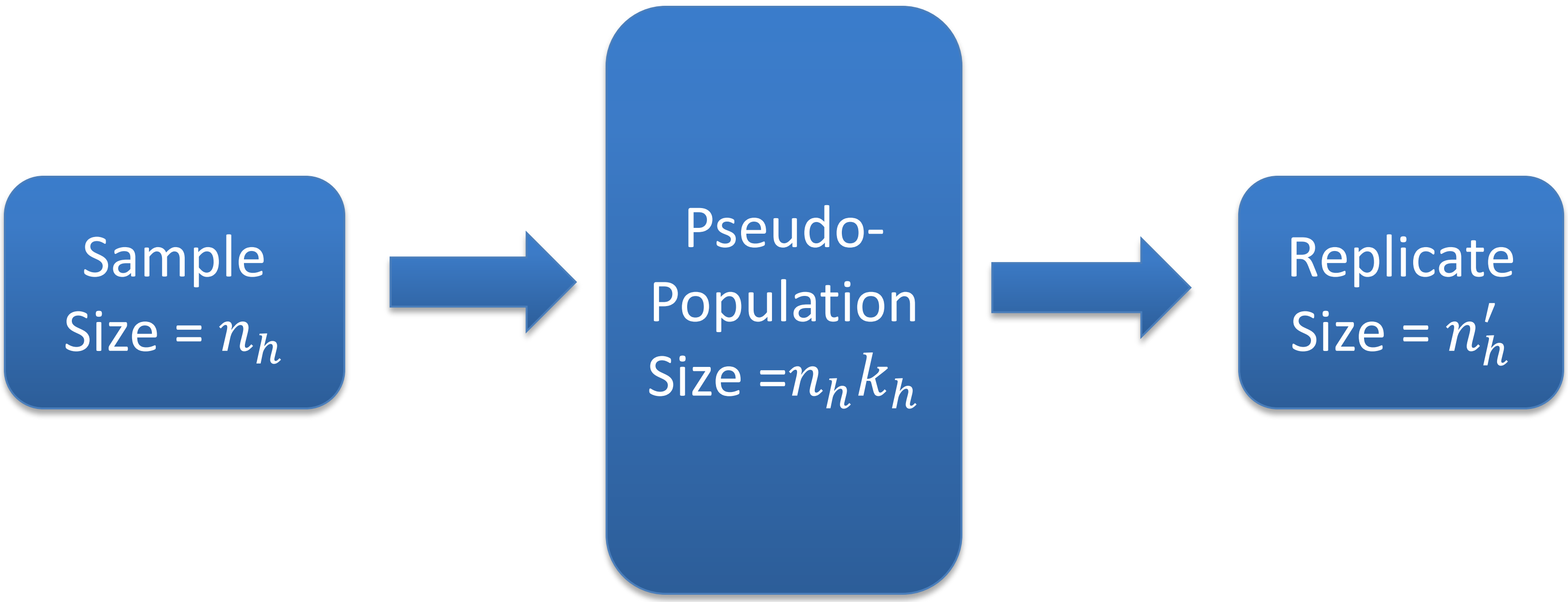
- Select a subsample of size n_h'

$$n_h' = f_h n_h$$

- Return subsample and repeat k_h times

$$k_h = n_h \left(1 - \frac{n_h'}{n_h}\right) / (n_h'(1 - f_h)$$

Without Replacement Bootstrap (BWO)



Without Replacement Bootstrap cont'd

- Create a pseudo-population by replicating each establishment k_h times

$$k_h = \frac{1}{f_h}\left(1 - \frac{1 - f_h}{n_h}\right)$$

- Create the replicate by selecting n'_h establishments from the pseudo-population

$$n'_h = n_h - (1 - f_h)$$

Creating Replicate Estimates

- Replicate estimates can be calculated as

$$\hat{Y}_{m,HT}^{(r)} = \sum_{h=1}^{H} \sum_{i=1}^{n'_h} w_{hi} y_i$$

- Ratio estimates can be calculated as

$$\hat{Y}_{m,ratio}^{(r)} = \sum_{h=1}^{H} \frac{RCPT_h}{\widehat{RCPT}_h} \sum_{i=1}^{n'_h} w_{hi} y_i$$

Creating Variance Estimates

The resulting estimate of variance is

$$v_{m,t} = C^{-1} \sum_{r=1}^{R} \left(\hat{Y}_{m,t}^{(r)} - \hat{Y}^{*} \right)^{2}$$

Where C $= \begin{cases} R & for\ MM\ and\ BWO \\ R-1 & for\ Chipperfield \end{cases}$

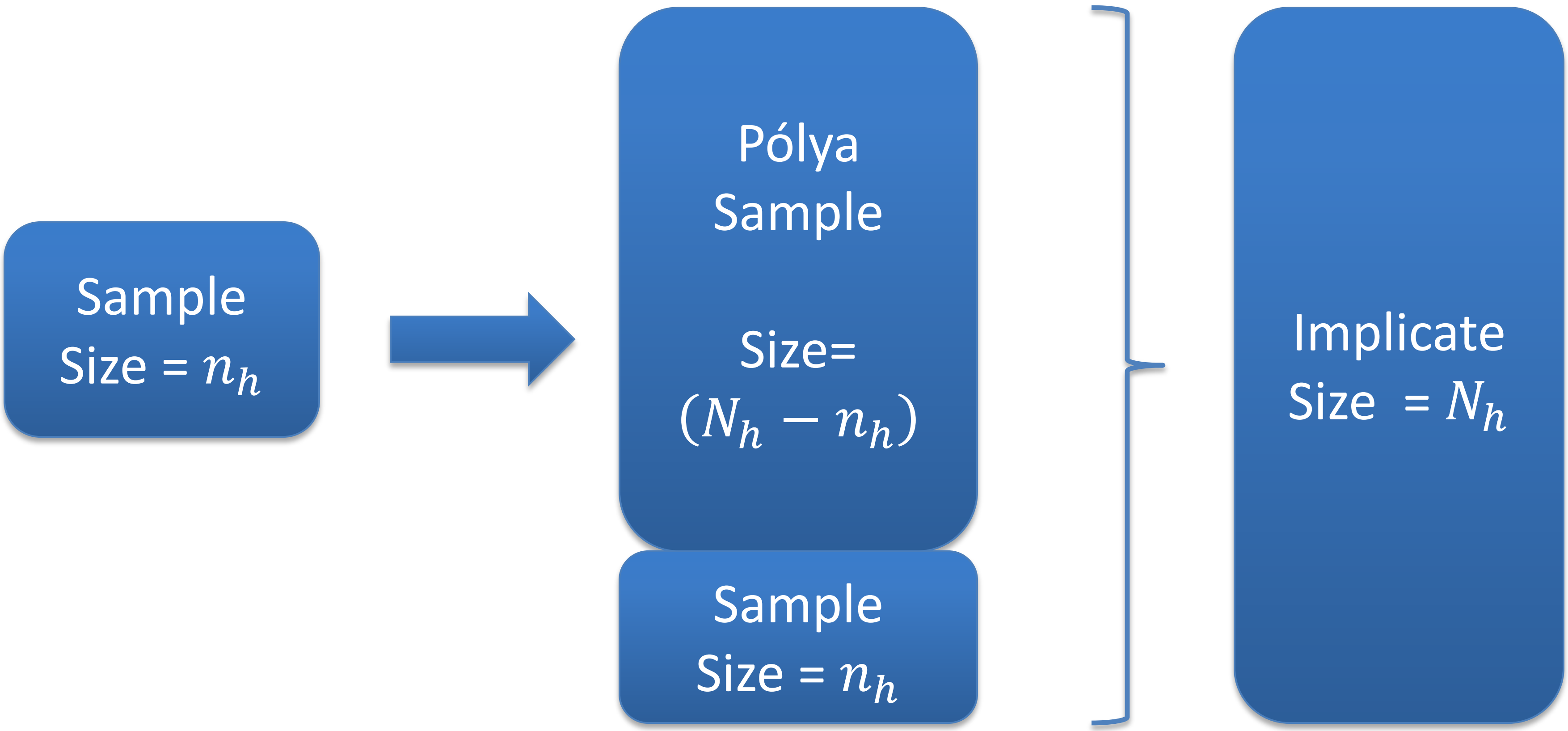
Finite Population Bayesian Bootstrap (FPBB)

- Create an implicate by drawing $N_h - n_h$ establishments from the sample with probability for the kth selection

$$p_{h,k} = \frac{\left(w_i - 1 + \frac{l_{i,k-1}(N_h - n_h)}{n_h}\right)}{N_h - n_h + \frac{(k_h - 1)(N_h - n_h)}{n_h}}$$

- Add the $N_h - n_h$ selected establishments to the original sample to complete the implicate

FPBB cont'd


Sample
Size = n_h
Pólya
Sample
Size=
$(N_h - n_h)$
Sample
Size = n_h
Implicate
Size = N_h

FPBB cont'd

The FPBB estimate of variance is

$$v_{FPBB} = U_B + \left(1 + \frac{1}{B}\right) T_B$$

where

- U_B is the average within-implicate variance
- T_B is the between-implicate variance, and
- B is the number of implicates

Evaluation

1. Full sample estimate versus mean of the replicate estimates
2. Number of replicates/implicates
3. Comparison of design-based methods
4. Comparison of selected design-based method to FPBB

Full Sample vs. Replicate Mean

		CHIP	MM	BWO
Horvitz-Thompson	Receipts	μ	μ	μ
	Product 1	μ	μ	
	Product 2	θ_0	θ_0	-
	Product 3	θ_0	θ_0	-
	Product 4	μ	μ	μ
Ratio	Product 1	θ_0	θ_0	-
	Product 2	θ_0	θ_0	-
	Product 3	θ_0	θ_0	-
	Product 4	μ	μ	μ

Evaluation cont'd

- The simulation study is a complete block design with repeated measures, treating industry as a random effect

- Used SAS PROC MIXED to fit and evaluate the following model

$$Y_{ij}^{l} = \tau_l + \beta_k + \gamma_{lk} + \epsilon_{ij}$$

$$\epsilon_i \sim N(0, \sigma^2), \beta_k \sim N(\beta, \sigma_\beta^2)$$

Number of Replicates/Implicates

100 vs 200 Replicates (CHIP, MM, BWO) — 10 vs 20 Implicates (FPBB)

	CHIP	MM	BWO	FPBB
P-values for Absolute Relative Bias (ARB)				
Product 1	1.00	1.00	1.00	0.92
Product 2	1.00	0.99	1.00	0.87
Product 3	1.00	1.00	1.00	0.04**
Product 4	1.00	0.99	1.00	0.87

P-values for Stability				
Product 1	1.00	1.00	0.97	0.62
Product 2	0.99	1.00	0.98	0.79
Product 3	1.00	1.00	0.99	0.00**
Product 4	0.99	1.00	0.99	0.88

Comparing Design-Based Methods

Estimator	Measure	Variable	Omnibus	BWO-CHIP	BWO-MM	CHIP-MM	Minimum Average
HT	ARB	Receipts	0.06**	0.04**	0.05**	0.95	BWO
HT	Stability	Receipts	0.01**	0.00**	0.08**	0.12	BWO
HT	Stability	Prod 4	0.08**	0.10**	0.03**	0.59	BWO
Ratio	Stability	Prod 4	0.08**	0.10**	0.03**	0.59	BWO

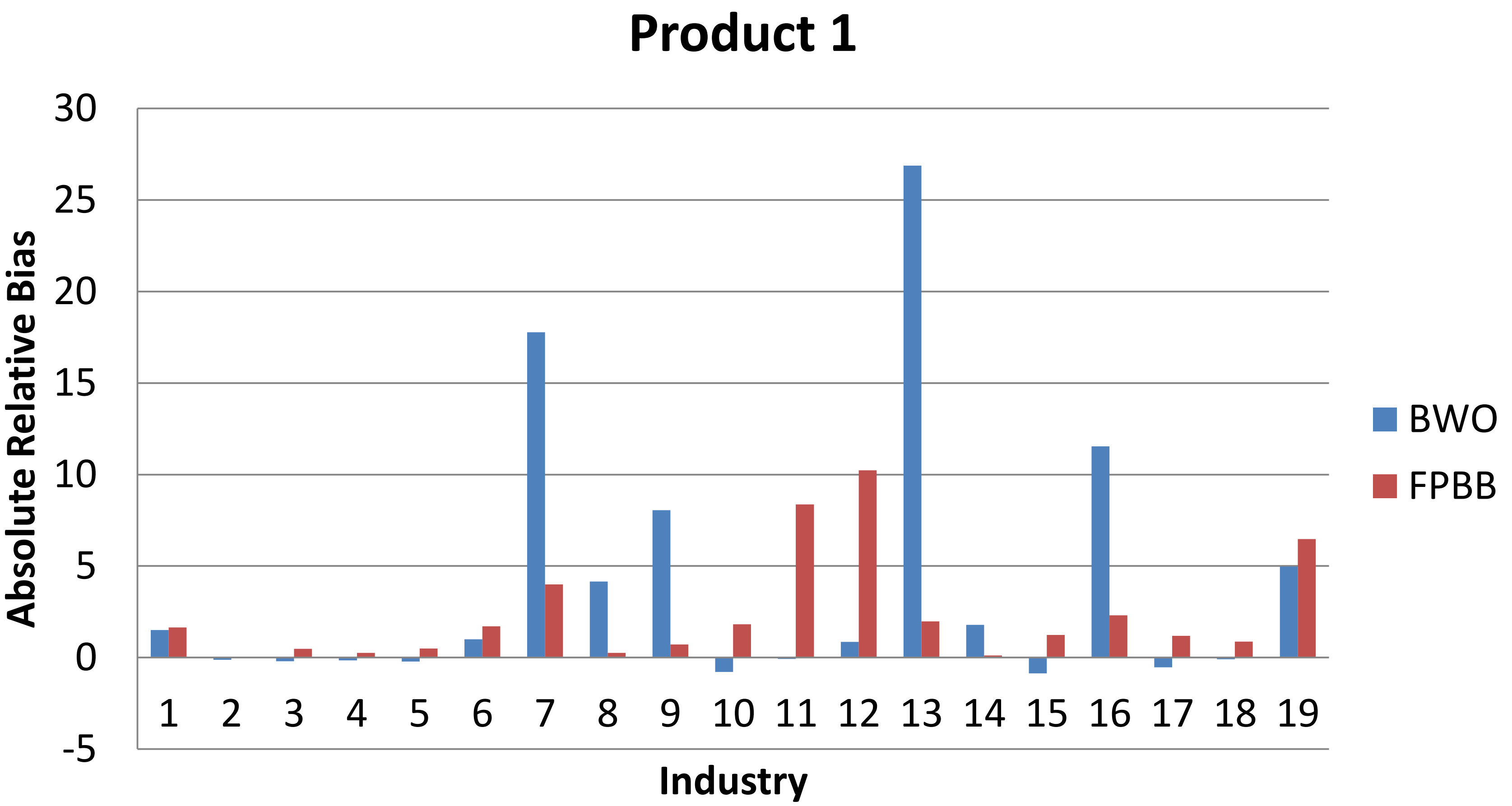
- Evaluated a total of 18 estimates (10 HT, and 8 Ratio)
- Only found significant differences in 4
- Minimal evidence of difference across methods

BWO vs. FPBB

- No significant differences identified

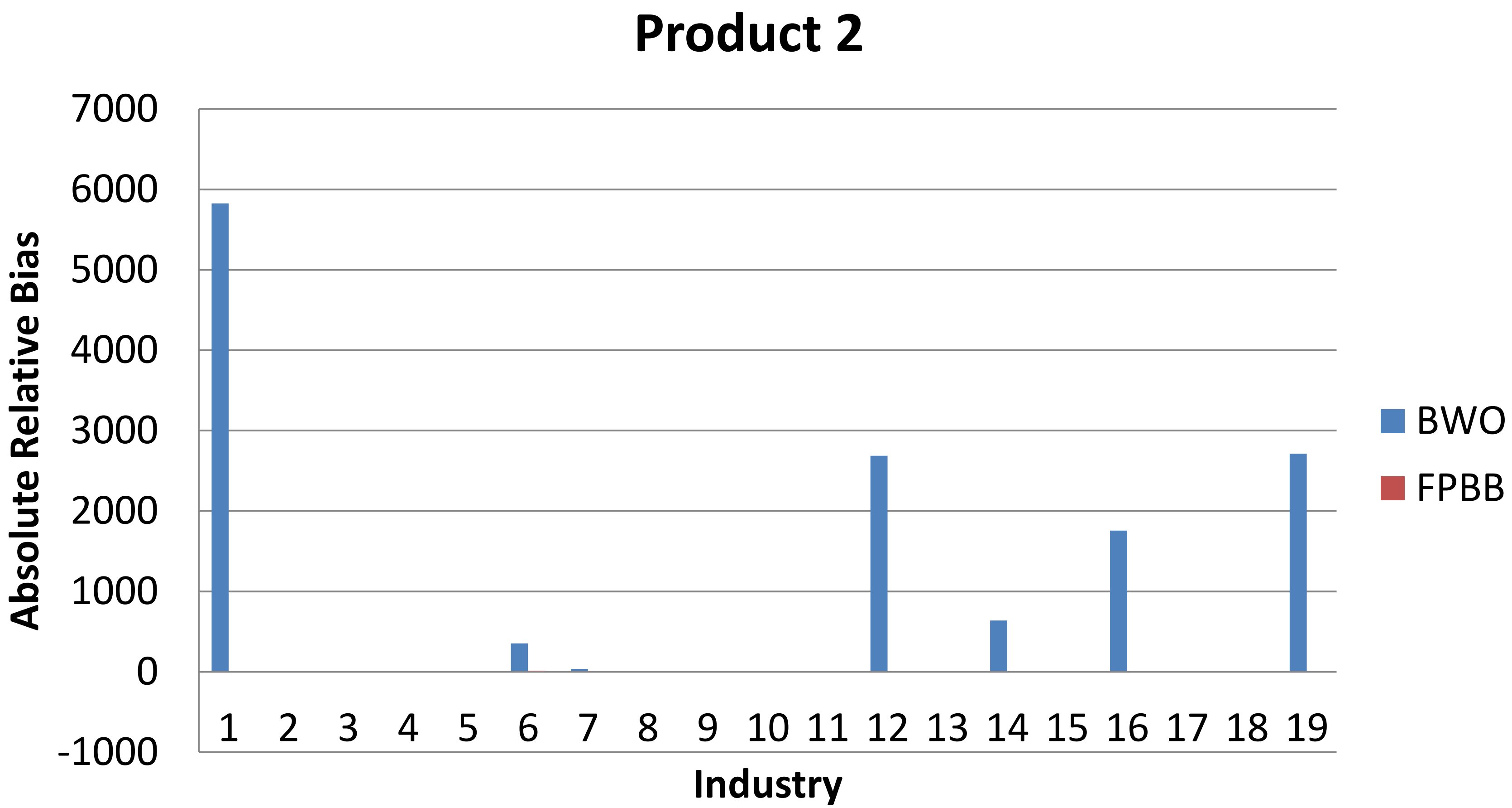
Ratio Estimates	ARB P-values	Stability P-values
Product 1	0.34	0.40
Product 2	0.34	0.37
Product 3	0.32	0.38
Product 4	0.34	0.38

BWO vs FBPP



Product 1
30
25
20
15
10
5
0
-5
Absolute Relative Bias
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Industry
BWO
FPBB

BWO vs FPBB



Product 2
7000
6000
5000
4000
3000
2000
1000
0
-1000
Absolute Relative Bias
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Industry
BWO
FPBB

Conclusions

In general, most establishments in an industry report the same products. The others are solicited but rarely reported.

In the first case, direct estimation is possible and the FPBB estimation method is feasible and preferable to the design-based replication methods investigated.

In the second case, the items are really small area estimates.

Next Steps

We will incorporate product nonresponse into the variance estimates by testing the FPBB/ABB method outlined in Zhou et al (2012) using a simulation approach combining the methods presented here with those presented in the previous presentation.

Acknowledgments

Robin Kurec, Laura James, Mark Pulling, Courtney Harris, Jeremy Knutson,

Victoria Garcia Toutsis, Glenn Zhen,

Rachel von Bargen

Contact Information

Matthew.Thompson@census.gov

Thank you!