Estimating Variances for the U.S. Census Bureau's Annual Integrated Economic Survey: Challenges in Estimation with a Low Entropy Sample Design

Katherine Jenny Thompson, Senior Mathematical Statistician

Associate Directorate for Economic Programs

Acknowledgements

Colt Viehdorfer*
Nicole Czaplicki
Stephen Kaputa
Mohammed Kashany
Matthew Thompson
Yeng Xiong

*Co-lead, with myself (katherine.j.thompson@census.gov)

This work builds on earlier research conducted by a team led by Justin Z. Smith.

Outline

- Introduce the AIES sample design and estimation procedures
- Present the candidate variance estimation procedures
  - Approximate sampling formula variance (SYG)
  - Antal Tillé Bootstrap (original)
  - Doubled Half Bootstrap
- Discuss our case study and resultant recommendation

Preliminaries

Annual Integrated Economic Survey (AIES)

- Economy wide survey conducted by U.S. Census Bureau
  - Mailout started in March 2024
  - Replaces seven annual surveys with one annual survey

- Publication Requirements
  - National industry estimates: **four** key items plus sector-specific items
    - C.V. target $\approx$ 2%
    - Disaggregated industry (NAICS) levels: NAICS3, NAICS4, **NAICS6**

  - Subnational (industry *x* geography) estimates: **four** key items
    - C.V. target $\approx$ 15%
    - Aggregated industry (NAICS) levels: **NAICS3**

Definitions

- Sector: an area of the economy in which businesses share the same or related business activity, product, or service (www.investopedia.com)

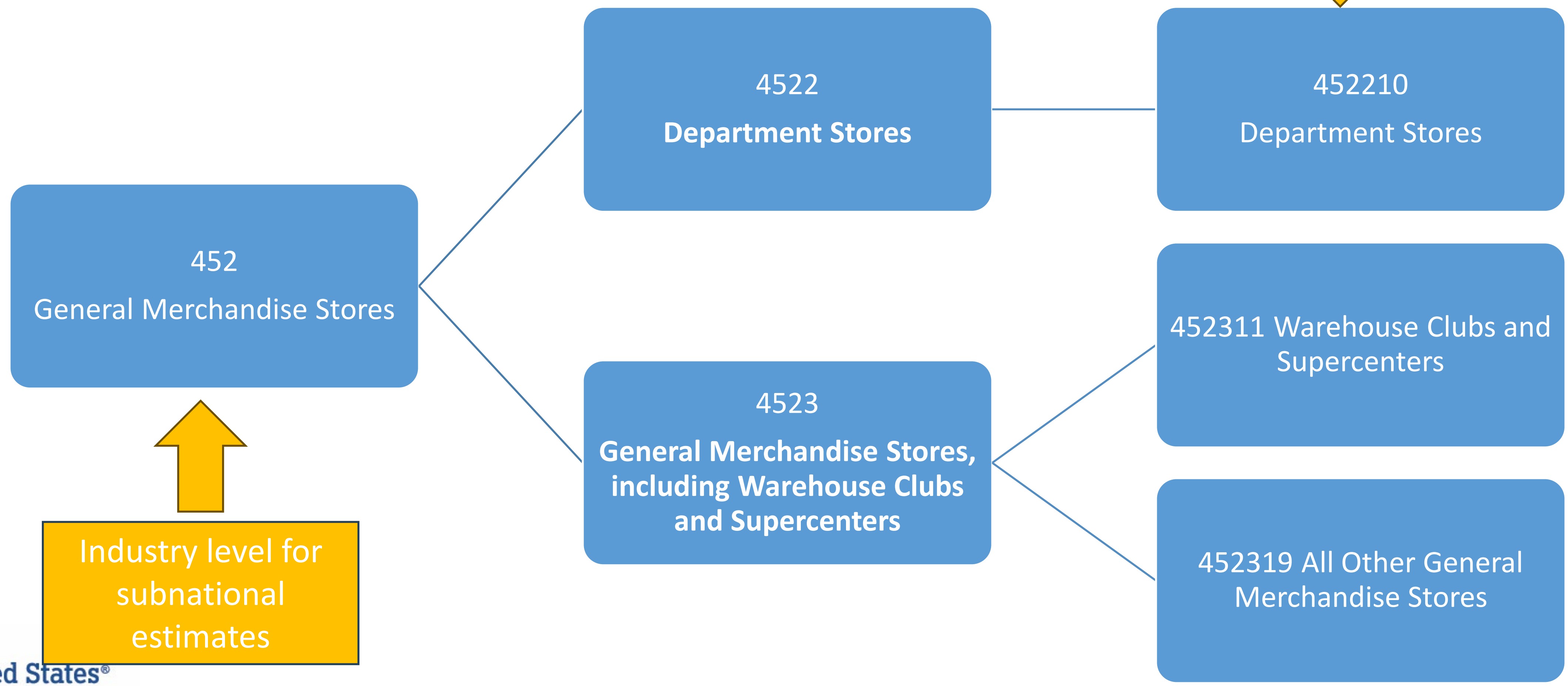
- Industry: a group of companies that are related based on their primary business activities or service (www.investopedia.com)

- Industrial classification: industry code assigned to an individual business, usually based on the business' largest source(s) of revenue

- NAICS: North American Industry Classification System
    - Digits indicate level of detail used for classification (more digits = more criteria)

United States® Census Bureau

Simple NAICS Example from the Retail Trade Sector



More Definitions

- Establishment – a single business location

- Company – any formal business entity for profit, which may be orati partnership, association or individual proprietorship (https://c y.l
  - Multi-unit company
  - pany

United States® Census Bureau

Sampling Frame...V

Sampling

Sampling Unit Components

Classification Information

Geo Infor

Quantitative Measure of Size (MOS)

Company	Establishment	ector	Industry	[redacted]	[redacted]	[redacted]	[redacted]
Global Dynamics, Inc.	Mom & Pop Hardware 1	Retail Trade	Hardware Stores	444130	MO	600	1,380
	Mom & Pop Hardware 2	[redacted]	[redacted]	444130	MO	800	[redacted]
	Mom & Po Hardware		Hardware Stores	444130	MO	762	1,753
	GD Plant	fig	B R Ma	332722	TN	9,452	[redacted]
	Dad's Warehouse	Wholesale Trade	Hardware Merchant Wholesalers	423710	TN	475	489
The Shop	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Multi-unit
Multi-sector
Multi-state

Single-units
Single-sector
Single-state

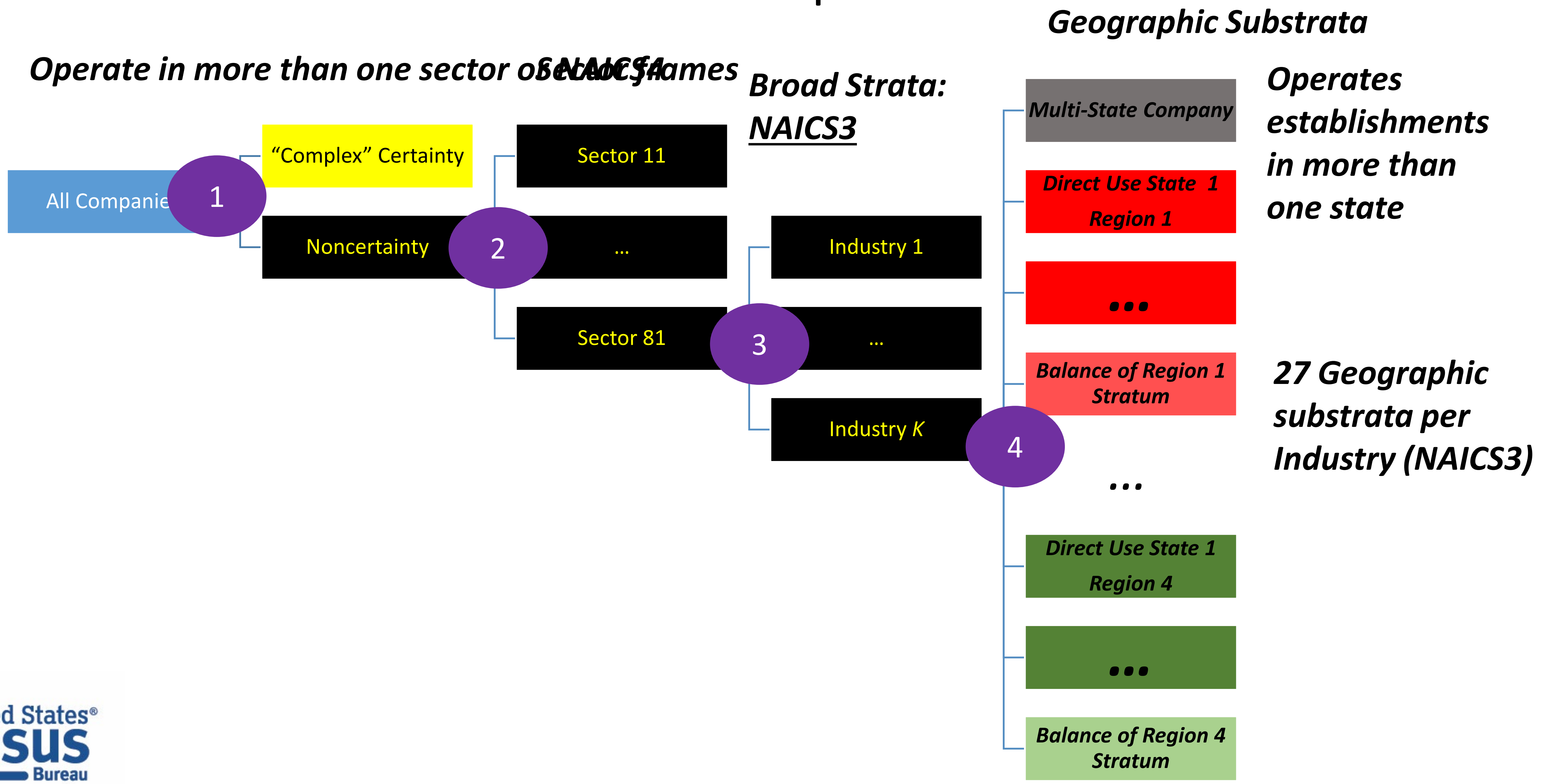
United States Census Bureau

AIES Sample Design

AIES Sample Design

- Sampling unit
  - Company
- Stratification
  - Subnational estimates within NAICS3
- Allocation
  - National and subnational estimates within NAICS3
  - Original allocations modified to achieve national NAICS6 C.V. targets
- Inclusion probability for company (π)
  - Accounts for company's contribution to national or subnational estimates (whichever is larger)

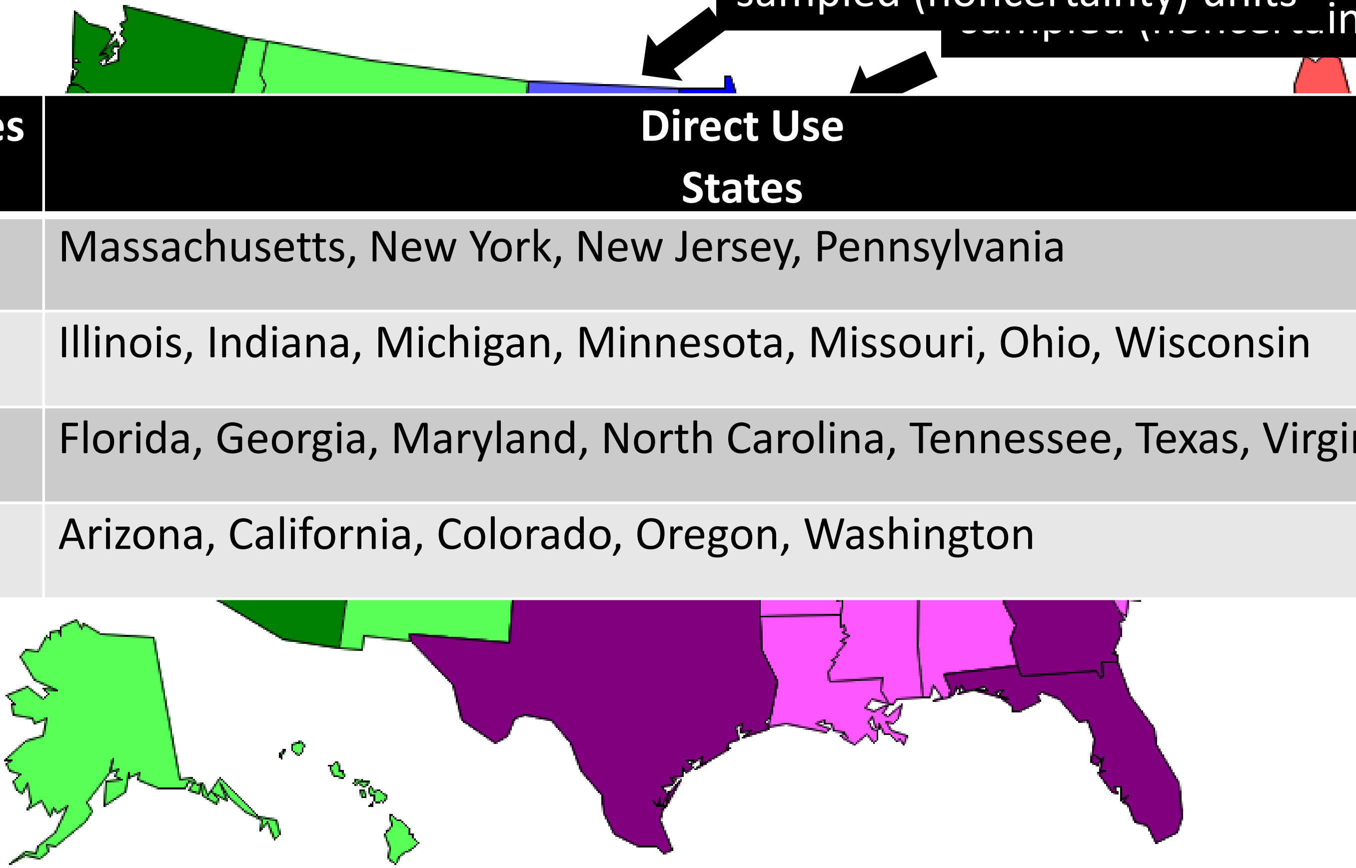
Basic Stratification of Companies



The 27 Geographic Substrata

Balance of Region Stratum:
Minimum number of
sampled (noncertainty) units

of
inty) units



Region	Total States	Direct Use States	Balance of Region States
Northeast	9	Massachusetts, New York, New Jersey, Pennsylvania	5
Midwest	12	Illinois, Indiana, Michigan, Minnesota, Missouri, Ohio, Wisconsin	5
South	16	Florida, Georgia, Maryland, North Carolina, Tennessee, Texas, Virginia	9 + DC
West	13	Arizona, California, Colorado, Oregon, Washington	8

AIES Sample Selection

- Sample Design Conditions
  1. Fixed sample size
  2. Unequal probability sampling
  3. Stratified sample
- Want
  - Sampled companies from each NAICS4, NAICS5, & NAICS6 within each NAICS3
  - Sampled companies from each state (for "balance of region")
  - Variety of company sizes
- Decision: List sequential sampling procedure
  - Sequential Random Sampling (Chromy's method)

High Versus Low Entropy Sampling

High Entropy

- Large variability between samples
- Joint inclusion probabilities can be expressed in terms of 1st order inclusion probabilities
- Examples
  - Simple random sampling
  - Pareto sampling
  - ***Poisson sampling***

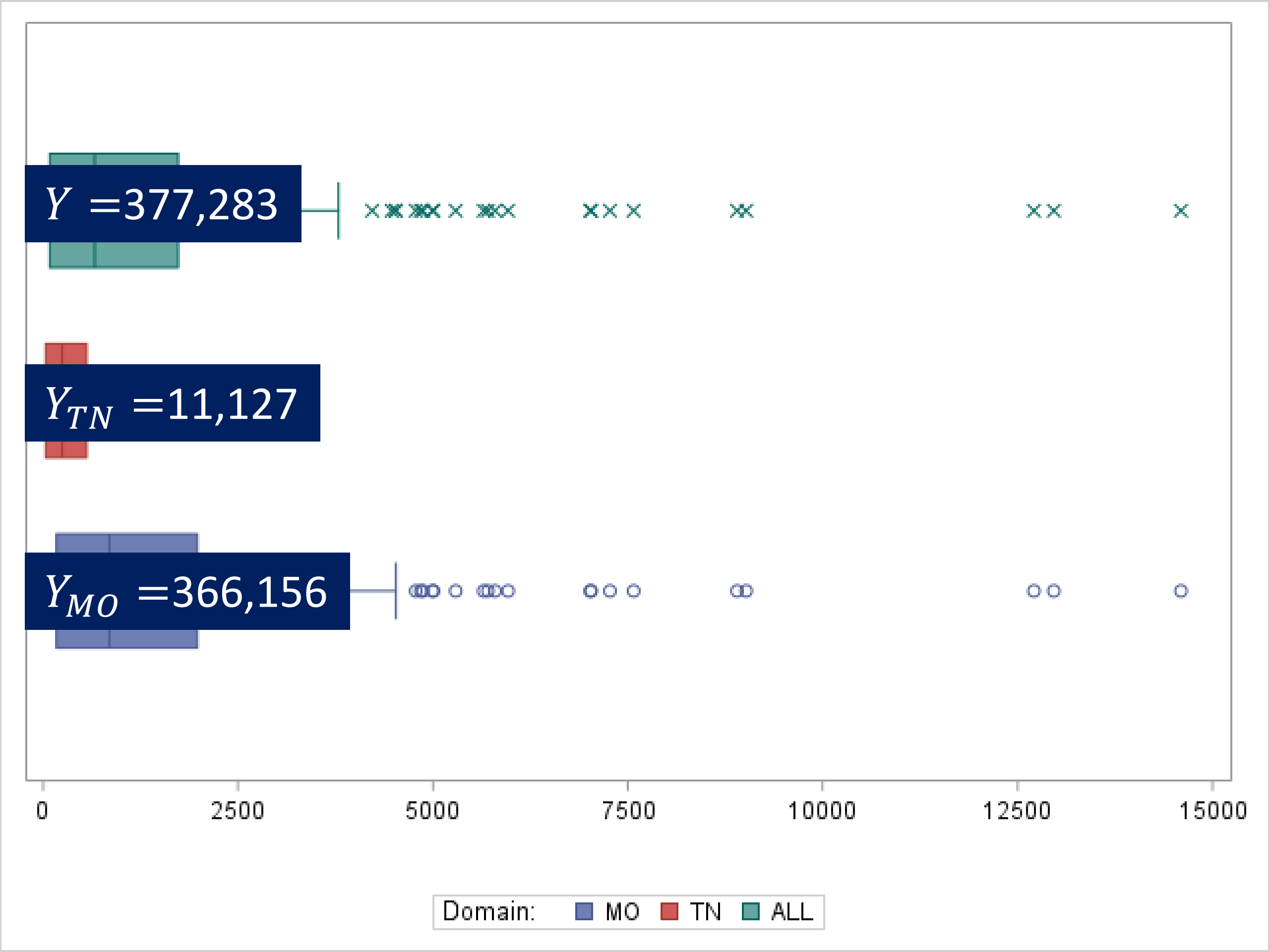
Low Entropy

- Little variability between majority of samples
- Joint inclusion probabilities **cannot** be expressed in terms of 1st order inclusion probabilities
- Examples
  - Systematic
  - Systematic PPS
  - ***Sequential Random (Chromy)***

Why Use (Low Entropy) Stratified Sequential Random Sampling?

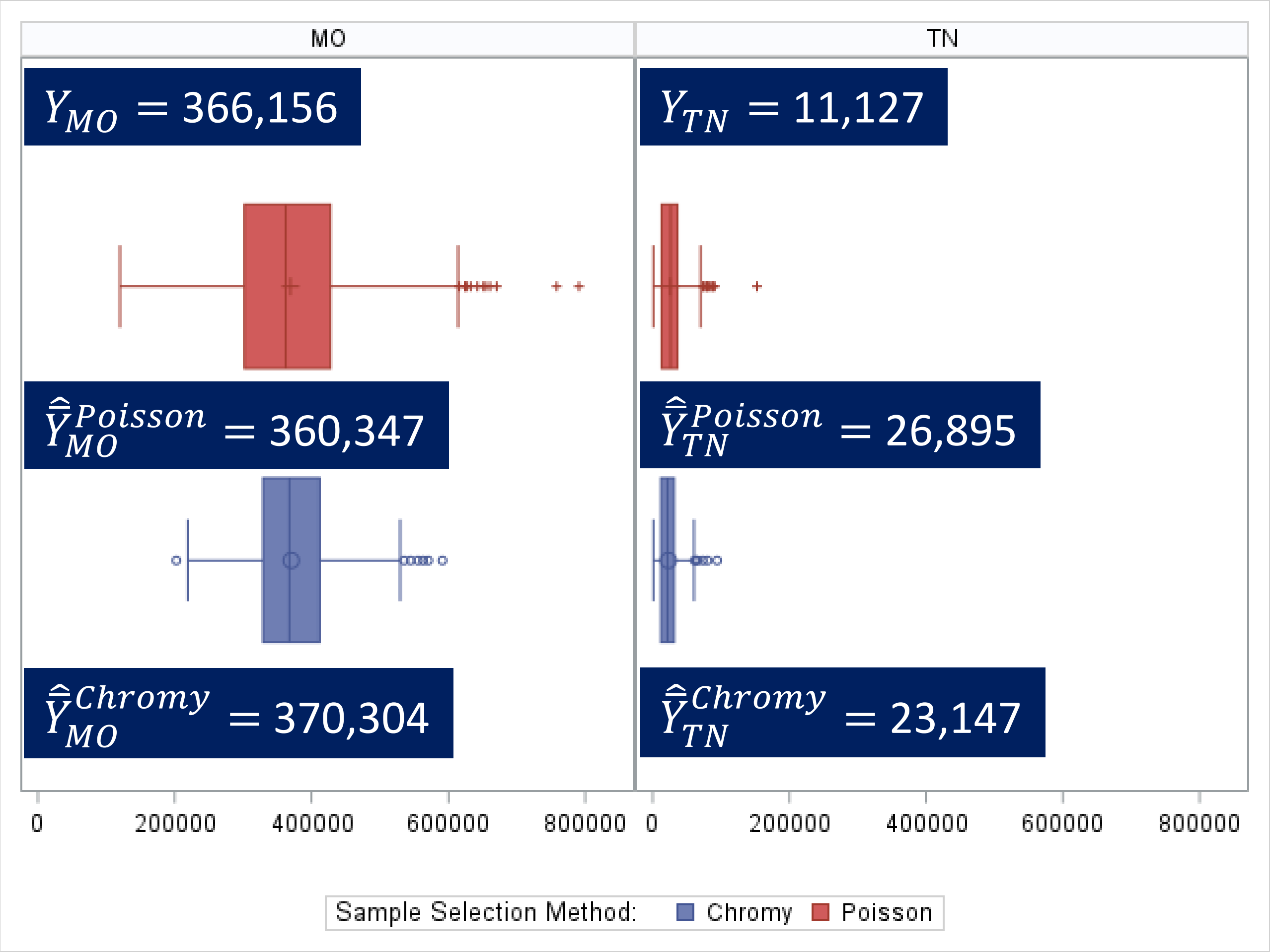
- Fictional business population
  - 280 companies
  - Sales (Y)
  - Payroll available (MOS)

- Two domains (states)
  - MO
  - TN



Mini-Simulation (Fictional Population)

- Select 1,000 random samples
  - *n* = 20 units
  - Sampling methods
    - Chromy
      - Sorted on state, payroll
    - Poisson
- Produce HT estimates of Sales
  - By State



United States® Census Bureau

AIES Estimation

- AIES produces DOMAIN estimates

- Horvitz Thompson (HT)

$$\hat{Y}^k = \sum_h \frac{y_{hi} I_{hi} I_{hi}^k}{\pi_{hi}} = \sum_h \frac{y_{hi}^k}{\pi_{hi}}$$

where y_{hi}^k = contribution to domain k estimate from company i

I_{hi} = 1 if company i in sampling stratum h is selected

$I_{hi}^k = 1$ if company i operates in domain k (0 otherwise)

- Ratio Estimator

$$\tilde{Y}^k = Y^{Ck} + \left(\frac{X^{NC,k}}{\hat{X}^{NC,k}}\right) \hat{Y}^{NC,k}$$

Frame total payroll (MOS) for domain k for variable y from noncertainty companies

Total contribution t

HT estimate for domain k for variable y from noncertainty (NC) companies $(\pi_{hi} < 1)$

HT estimate of payroll for domain k for variable y from noncertainty companies

AIES Candidate Variance Estimators

Sen Yates Grundy Variance Estimator
Horvitz Thompson Estimate

$$\hat{v}_{SYG}(\hat{Y}^k) = \sum_{h=1}^{H} \frac{1}{2} \sum_{i=1}^{n_h} \sum_{l \neq i} \left(\frac{\pi_{hi}\pi_{hl} - \pi_{h,il}}{\pi_{h,il}} \right) \left(\frac{y_{hi}^k}{\pi_{hi}} - \frac{y_{hl}^k}{\pi_{hl}} \right)^2$$

KNOWN inclusion probability for company *i* in stratum *h*

Joint inclusion probability of companies *i* and *l* in stratum *h*

- = 0 if companies are in different strata
- Dependent on <u>full sample</u> design (LOW entropy design)

Challenges with $\hat{v}_{SYG}(\hat{Y}^k)$

Methodology

- Unbiased estimate when $n_h \geq 2$
  - Not always true for AIES samples
  - Collapsed stratum estimator
- Sensitive to joint inclusion probabilities $(\pi_{h,il})$
  - Cannot be expressed as function of first order inclusion probability with Chromy sampling
  - Estimation of $(\pi_{h,il})$
    - Exact for $n_h \leq 3{,}000$
    - Resampling approximation otherwise

Implementation

- Difficult to store $\pi_{h,il}$ for all company pairs in database
- NUMEROUS obstacles when $y_{hi}^k = 0$

Replication method

Replication Methods Presented

- Bootstrap resampling methods
  - Same size as original sample
  - Same design weights as original sample

- Resampling is <u>different</u> from original sampling scheme
  - Incorporates finite population correction
  - Approximates design-based variance estimator for unequal probability samples

- Evaluated on high entropy sampling designs
  - Theory developed with PPS sampling
  - AIES implements PP<u>E</u>S sampling

Replication Methods

Antal Tillé Bootstrap (ATB)

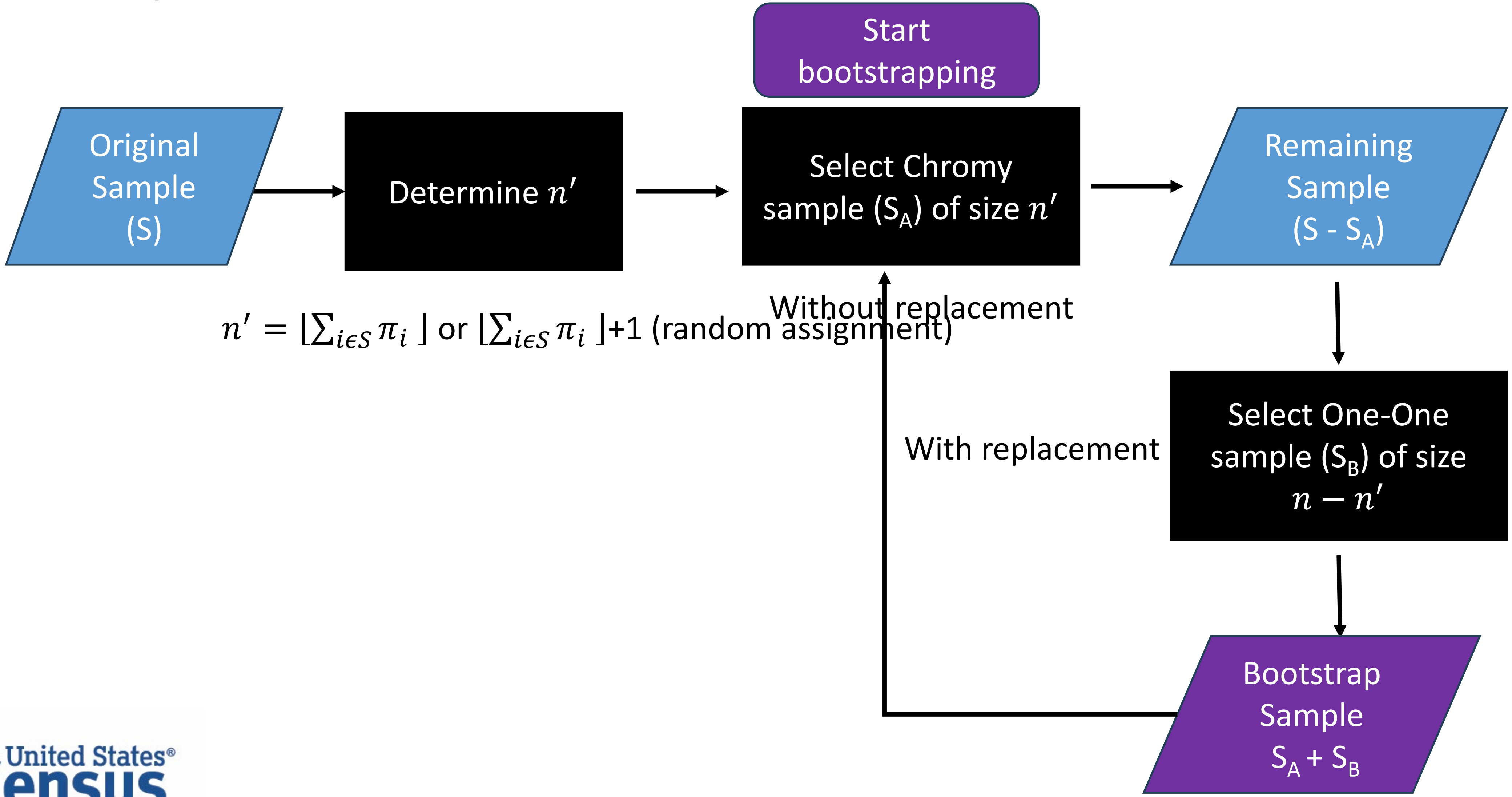
- Sampling procedure (mixed)
  - Without replacement using parent sampling method
  - With replacement
- Equals *approximation* to SYG variance estimator
  - Requires selecting one of three possible approximation of $\pi_{h,il}$ as function of π_{hi}

Doubled Half Bootstrap (DHB)

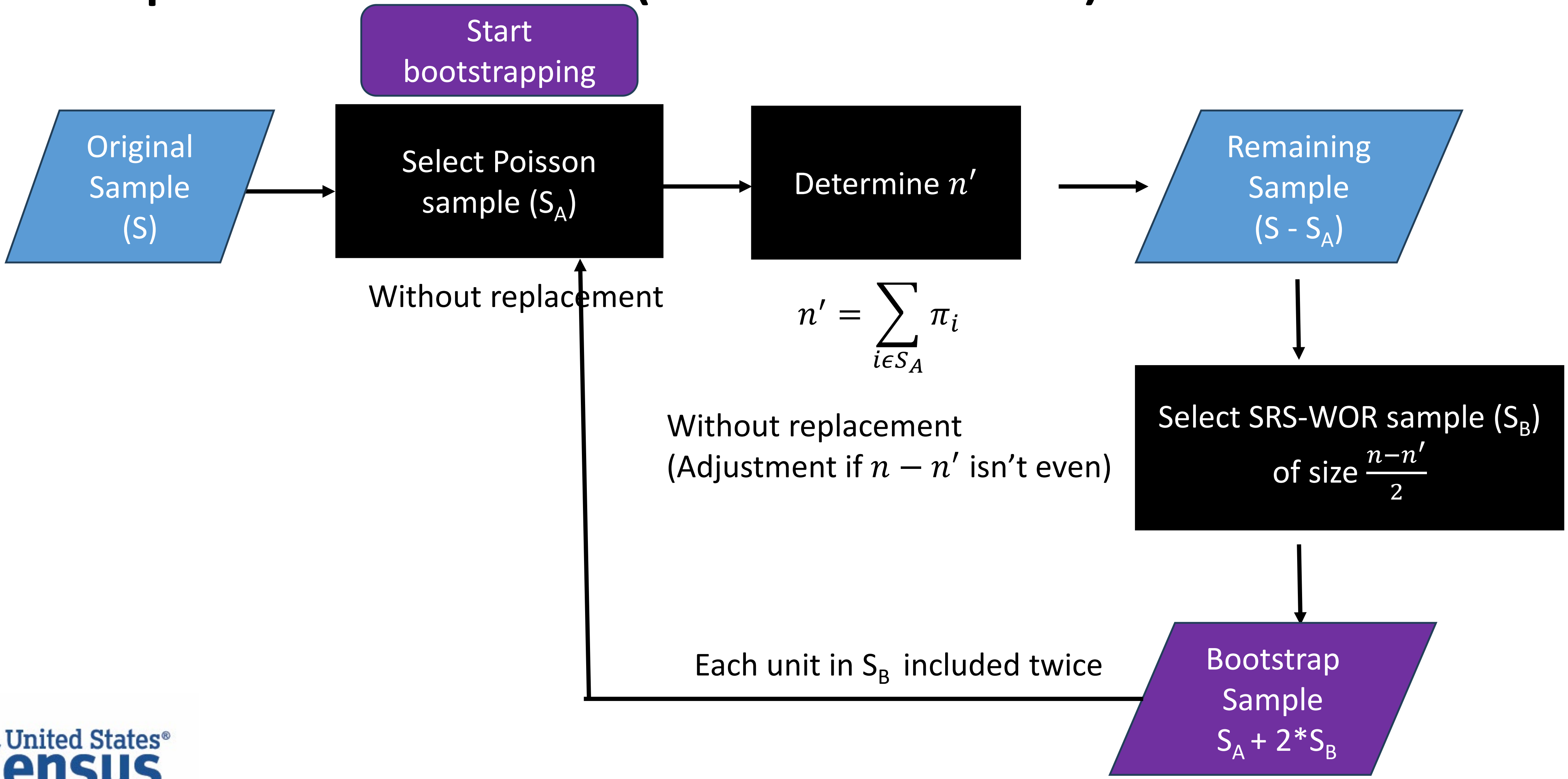
- Sampling procedure (mixed)
  - Without replacement (Poisson)
  - With replacement
- Equals *approximation* to SYG variance estimator
  - <u>One</u> approximation of $\pi_{h,il}$ as function of π_{hi} (no choice)

United States Census Bureau

Antal Tillé Bootstrap (ATB)
Simplified Picture (One Stratum)



Doubled Half Bootstrap (DHB) Simplified Picture (One Stratum)



Bootstrap Variance Estimator

$$v_m(\hat{\theta}) = \frac{1}{B-1}\sum_b \left(\hat{\theta}_m^b - \hat{\bar{\theta}}_m\right)^2$$

m = ATB or DHB

Case Study

Case Study

- Seven NAICS3 industries (Frame Data)
  - All sampling strata have $n_h \leq 3{,}000$ (can obtain exact $\pi_{h,ij}$)
  - All industries contain two or more NAICS6 categories

- Target domain estimates: NAICS3 *x* Direct-Use State
  - 23 per NAICS3
  - Horvitz-Thompson estimates
    - Ratio adjusted estimates “mask” differences in variance estimators

- Do not consider NAICS6 national estimates or NAICS3 by “balance of state” estimates in presented evaluation
  - Too much confounding due to random sample size

United States® Census Bureau

Case Study

- 5,000 independent Chromy samples per industry
  - AIES design strata, allocations, and 1st order inclusion probabilities
  - Construct empirical variance ($\mathrm{V}(Y^k)$)

- 1,000 of the 5,000 samples
  - SYG variance estimator (collapsed stratum)
  - ATB variance estimator
  - DHB variance estimator

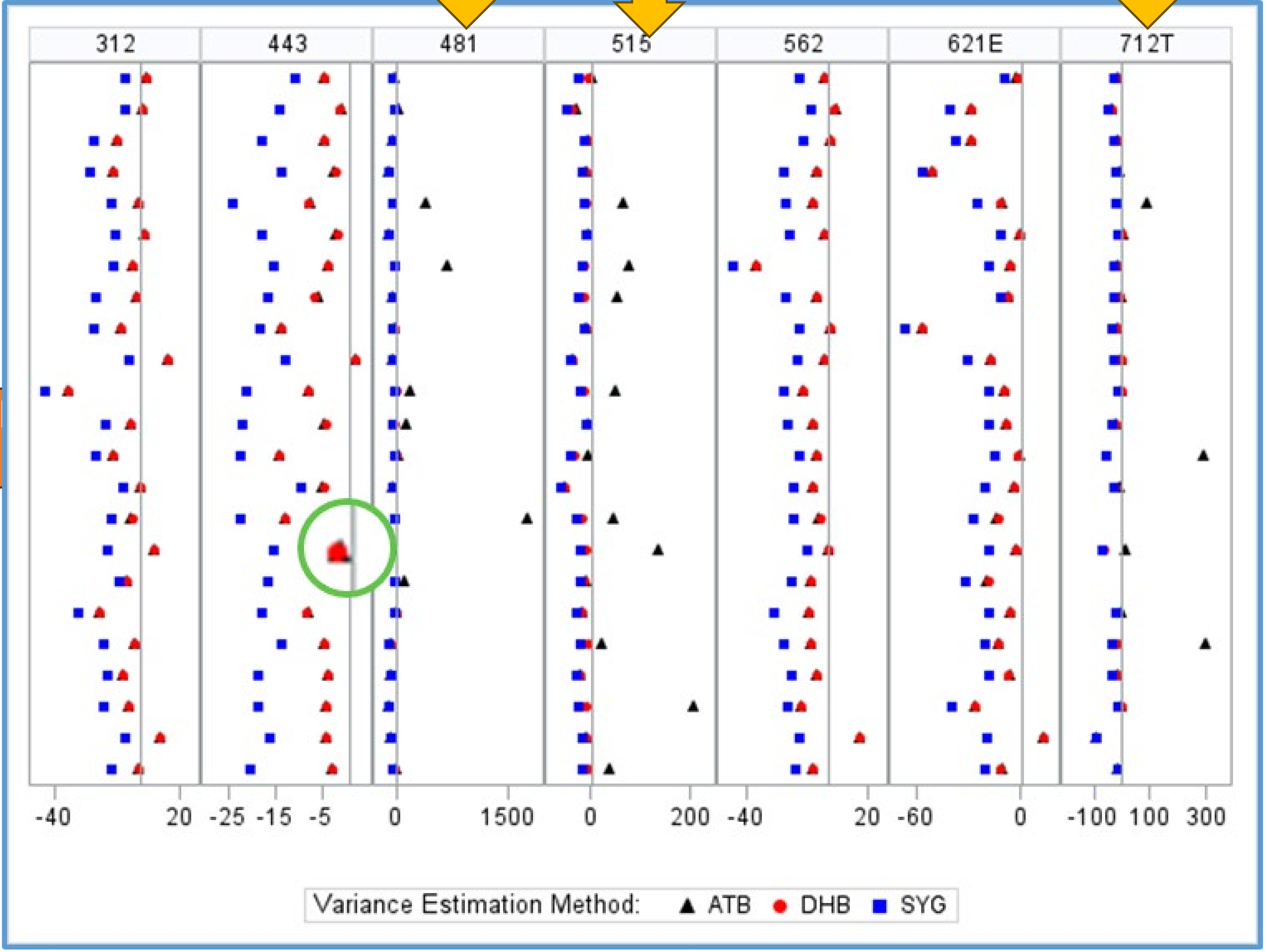
- 400 bootstrap replicates per sample

Evaluation Statistics

- Standard Error Ratio: $R_m^k = \dfrac{\frac{1}{1{,}000}\sum_s \sqrt{\hat{v}_{ms}\left(\hat{Y}^k\right)}}{\sqrt{V\left(\hat{Y}^k\right)}}$

- 90% confidence interval coverage rate

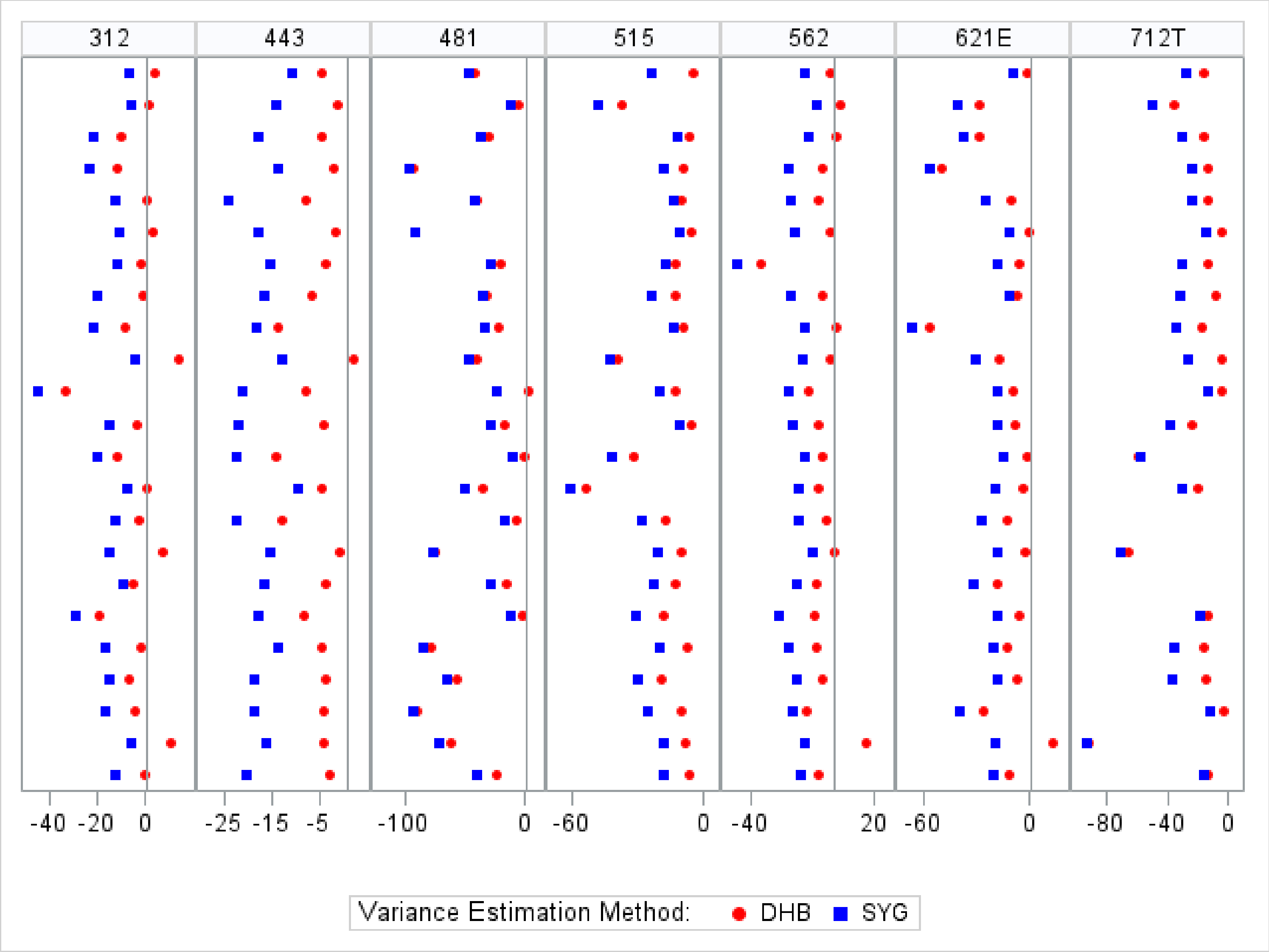
Standard Error Ratios (All)



Reference at 1

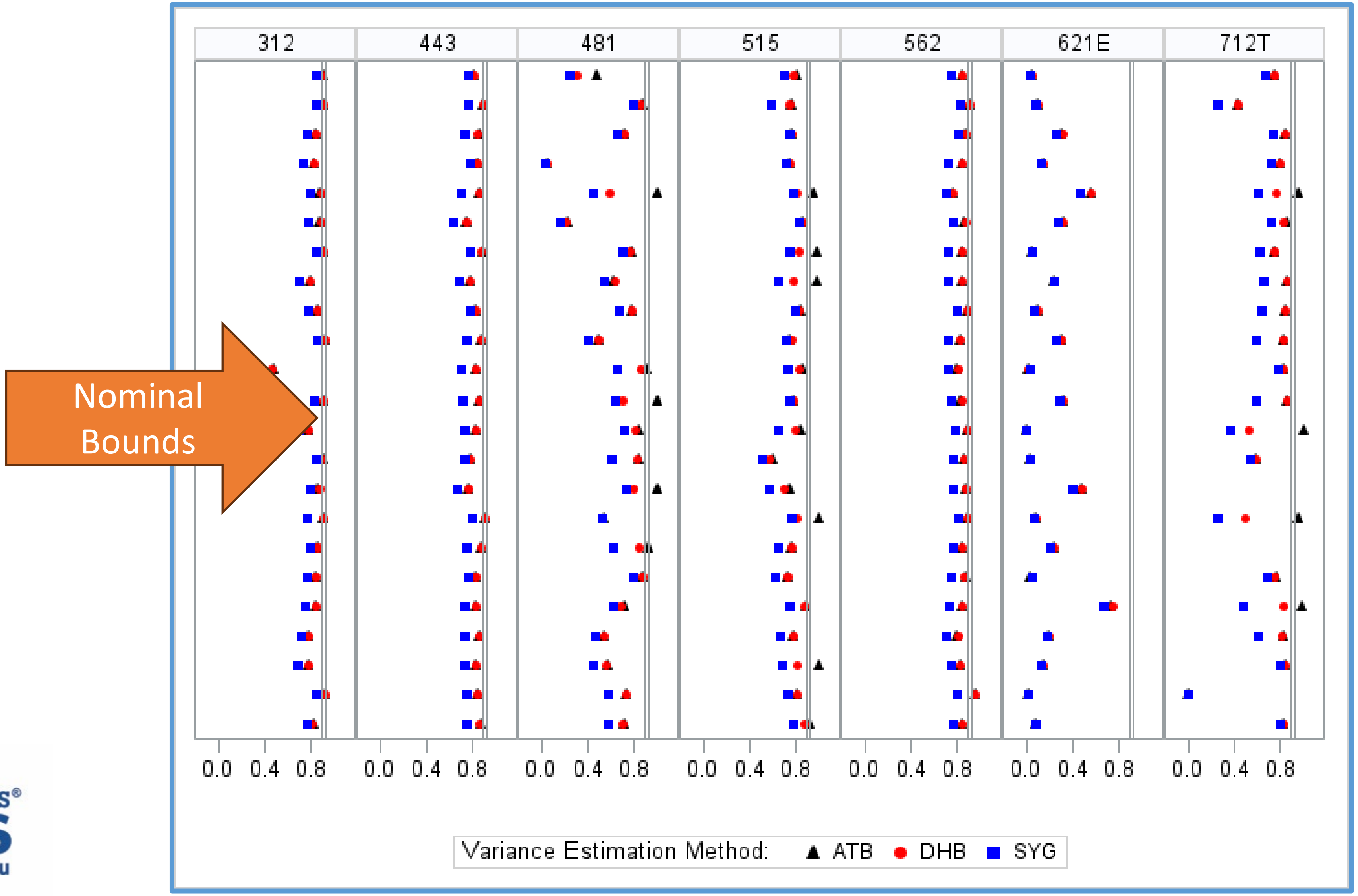
Project No. P-7529180, Disclosure Review Board (DRB) approval number: CBDRB-FY25-ESMD010-002

Standard Error Ratios (No ATB)



Project No. P-7529180, Disclosure Review Board (DRB) approval number: CBDRB-FY25-ESMD010-002

90% Confidence Interval Coverage Rates (All)



Project No. P-7529180, Disclosure Review Board (DRB) approval number: CBDRB-FY25-ESMD010-002

However, there is confounding!

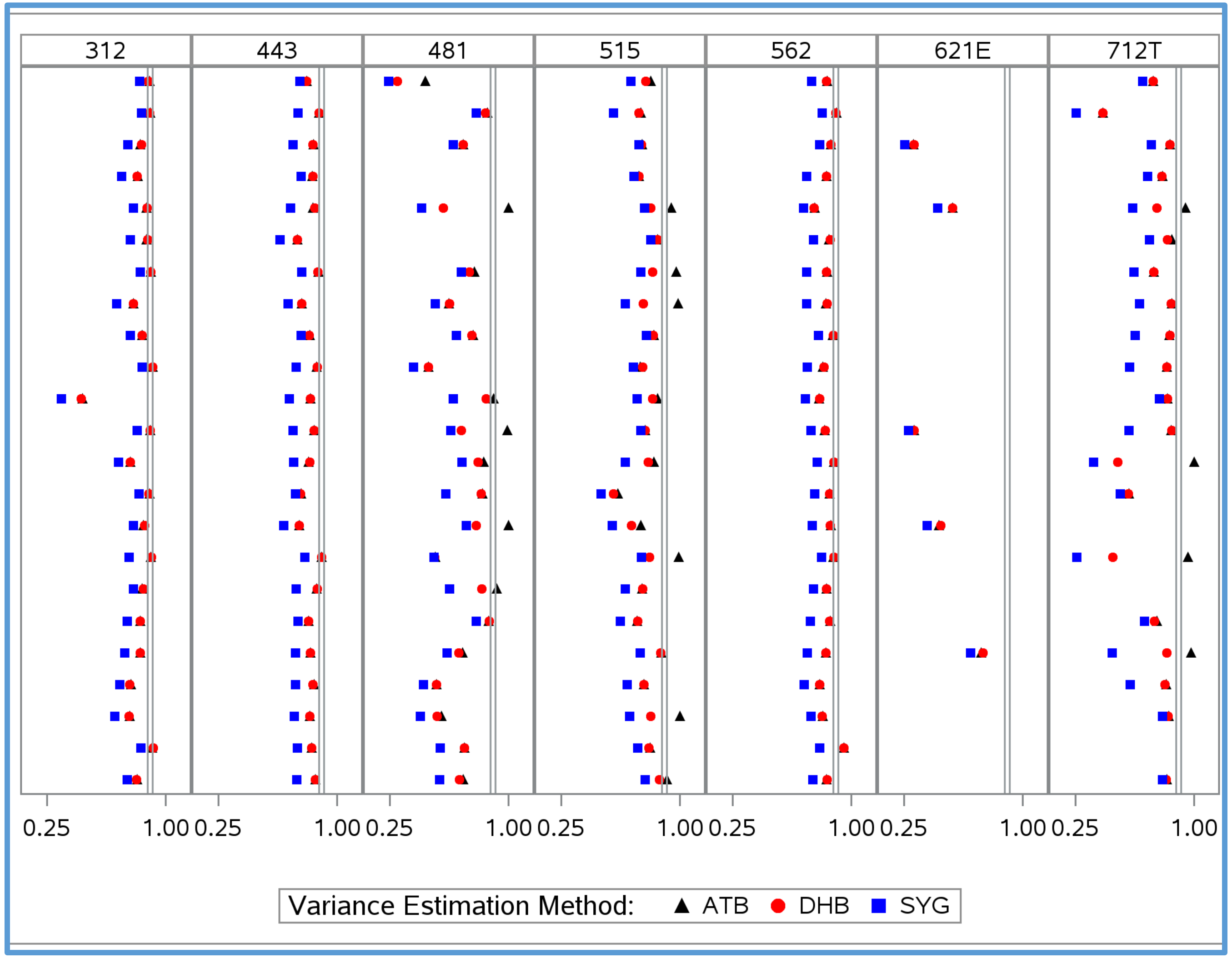
In our 1,000 independent samples,

$$\textit{ABS(Relative Bias (RB))} = \left| \left[\frac{\frac{1}{1{,}000} \sum_{s=1}^{1{,}000} \hat{Y}_s}{Y_s} \right] - 1 \right|$$

- Is less than 0.03 ("unbiased") for all state estimates in industries 312, 443, 515, & 562
- Is <u>greater</u> than 0.03 ("biased") in industries 481 and 712T for one state estimate apiece
- Is greater than 0.03 in industry 621E for 18 (of 23) state estimates

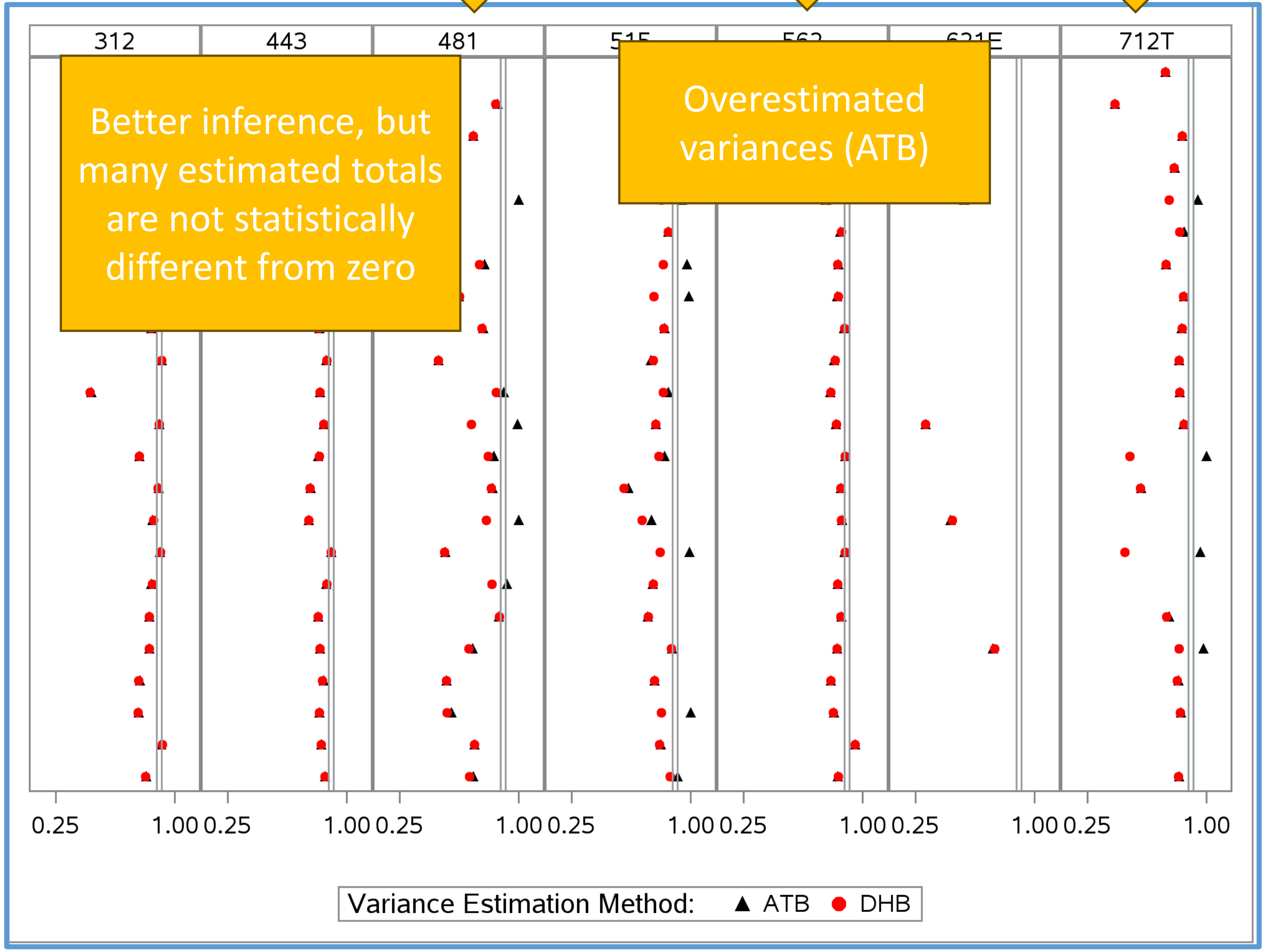
United States® Census Bureau

90% Confidence Interval Coverage Rates (All) Excluding Biased Estimates



Project No. P-7529180, Disclosure Review Board (DRB) approval number: CBDRB-FY25-ESMD010-002

90% Confidence Interval Coverage Rates



Project No. P-7529180, Disclosure Review Board (DRB) approval number: CBDRB-FY25-ESMD010-002

United States® Census Bureau

Conclusion

- Revelations
  - Severe undercoverage with SYG estimator
  - Decent performance (certainly better) with replication methods
    - ATB – inconsistent performance due to Chromy sampling
    - DHB – consistent performance
- Recommendation for AIES: doubled half bootstrap (DHB)
  - For production estimates
  - In future sample selection programs for AIES

United States® Census Bureau

References

Antal, E., & Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, *106*(494), 534–543.

Antal, E., & Tillé, Y. (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics*, 29, 1345-1363.

Chauvet, G. (2021). A note on Chromy's sampling procedure. *Journal of Survey Statistics and Methodology*, 9(5), 1050–1061.

Chromy, J. R. (1979). Sequential sample selection methods. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 401–406.

Questions, suggestions, comments?

katherine.j.thompson@census.gov

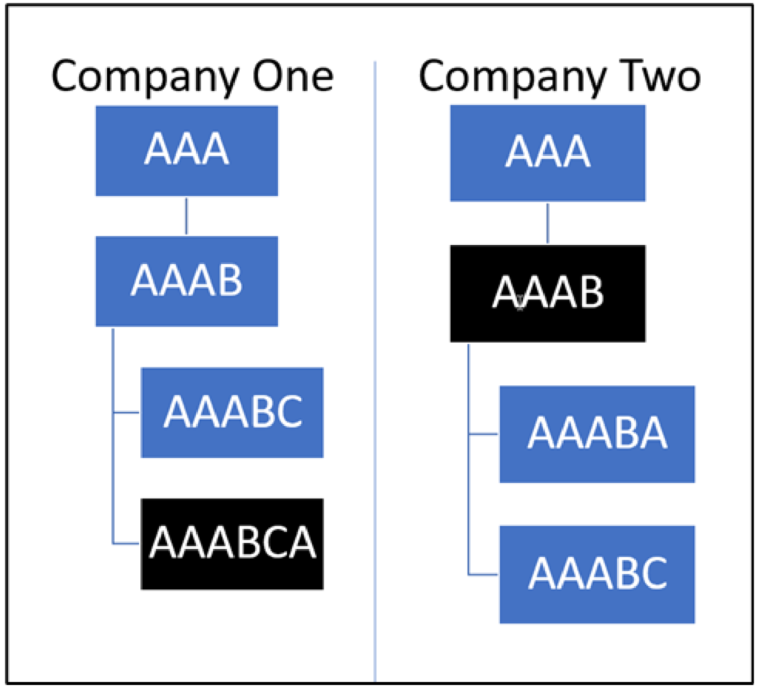
Extra Slides

Certainty Companies in AIES

- "Complexity"
  - Based on company economic diversity (2+ sectors, 2+ NAICS4)
- Allocation
  - 15 or fewer units in sampling stratum
  - More sample allocated than number of companies in stratum
- PPES
  - Function of company probability of selection and stratum sample size

Stratified Sequential Random Sampling for AIES

- Sort companies within each noncertainty stratum by
  - Lowest NAICS classification
  - State
  - Inclusion probability
- Select companies from ordered list



Sen Yates Grundy Variance Estimator Ratio Adjusted Estimate

$$\hat{v}\left(\tilde{Y}^k\right) = \left(\hat{R}_s^k\right)^2 \sum_{h=1}^{H} \frac{1}{2} \sum_{i=1}^{n_h} \left(\frac{\pi_{hi}\pi_{hl} - \pi_{h,il}}{\pi_{h,il}}\right) \left(\frac{\hat{e}_{hi}^k}{\pi_{hi}} - \frac{\hat{e}_{hi}^k}{\pi_{hl}}\right)^2$$

Ratio adjustment factor (noncertainty units)

Linearized ratio estimate

Bootstrap Variance Estimator

$$v_m\left(\hat{\theta}\right) = \frac{1}{B-1}\sum_b \left(\hat{\theta}_m^b - \hat{\bar{\theta}}_m\right)^2$$

m = ATB or DHB

Horvitz Thompson estimate from bootstrap replicate *b* (method *m*)

Average over all bootstrap replicates

B = number of bootstrap replicates