Candidate Sample Designs

	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	DOES NOT ACCOUNT FOR UNIT SIZE
PPS	Pareto sample without replacement	No stratification

United States® Census Bureau

Candidate Sample Designs

	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	No stratification
PPS	Pareto sample without replacement	ACCOUNTS FOR UNIT SIZE $\pi_{ij} = n_i \left(\frac{X_{ij}}{\sum_{j \epsilon i} X_{ij}} \right)$

United States® Census Bureau

Candidate Sample Designs

	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	No stratification
PPS	Pareto sample without replacement	No stratification
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	6 strata per industry

United States® Census Bureau

Candidate Sample Designs

	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	No stratification
PPS	Pareto sample without replacement	No stratification
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	

ACCOUNTS FOR UNIT SIZE
X (MOS) used to define strata boundaries

United States® Census Bureau

Candidate Sample Designs

	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	No stratification
PPS	Pareto sample without replacement	No stratification
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	6 strata per industry
SSRS_LH	Stratified SRS, Lavallée Hidiroglou (LH) method and Dalenius Hodges (DH)	1 certainty (take all) stratum 5 noncertainty strata

United States® Census Bureau

Candidate Sample Designs

	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	No stratification
PPS	Pareto sample without replacement	No stratification
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	Include "Certainty" (Take All Stratum)
SSRS_LH	Stratified SRS, Lavallée Hidiroglou (LH) method and Dalenius Hodges (DH)	atum

United States® Census Bureau

Example: Food Manufacturing

- BERD sample
  - Prevalence
    - ≈ 20% of companies have R&D expenditures (2018 and 2019 samples)
  - Current R&D Expenditures ≈
    - (0.0738)(Annual Payroll) (adj-R^2 =0.27)
    - (0.6990)(Prior R&D Expenditures) (adj-R^2 =0.95)

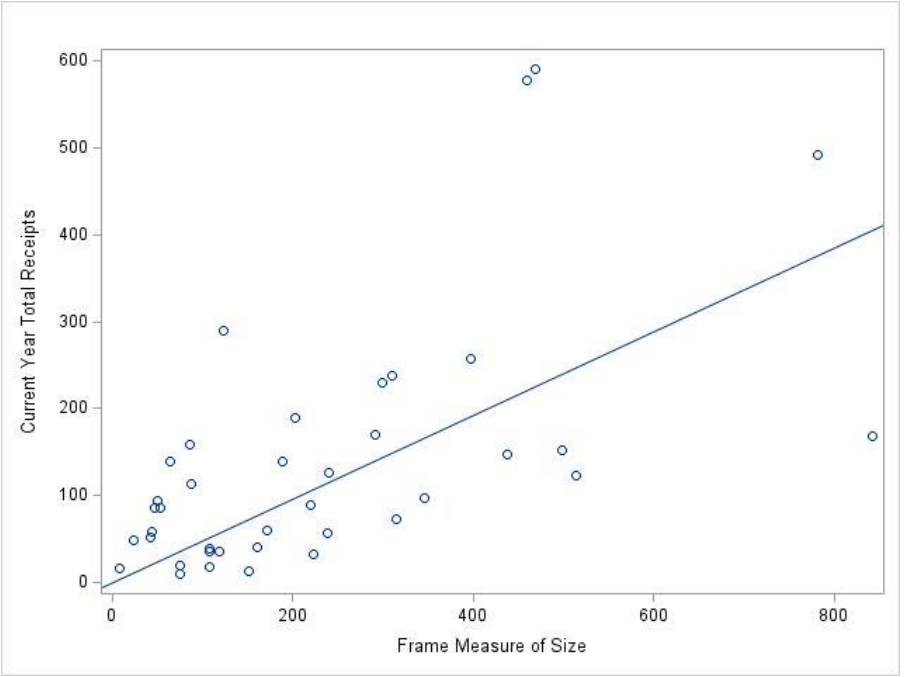
- Simulation study allocation = 1,400 companies
  - Proportional allocation
  - Sampling fraction = 0.1528

United States® Census Bureau

Recall“Typical” (Business) Survey Design Setting

- Sample design utilizes frame measure of size

- Frame variables used to evaluate effectiveness of design
  - One or more candidate sample designs
  - Assess performance by
    - Drawing one sample from single frame
    - Drawing repeated samples from single frame



United States® Census Bureau

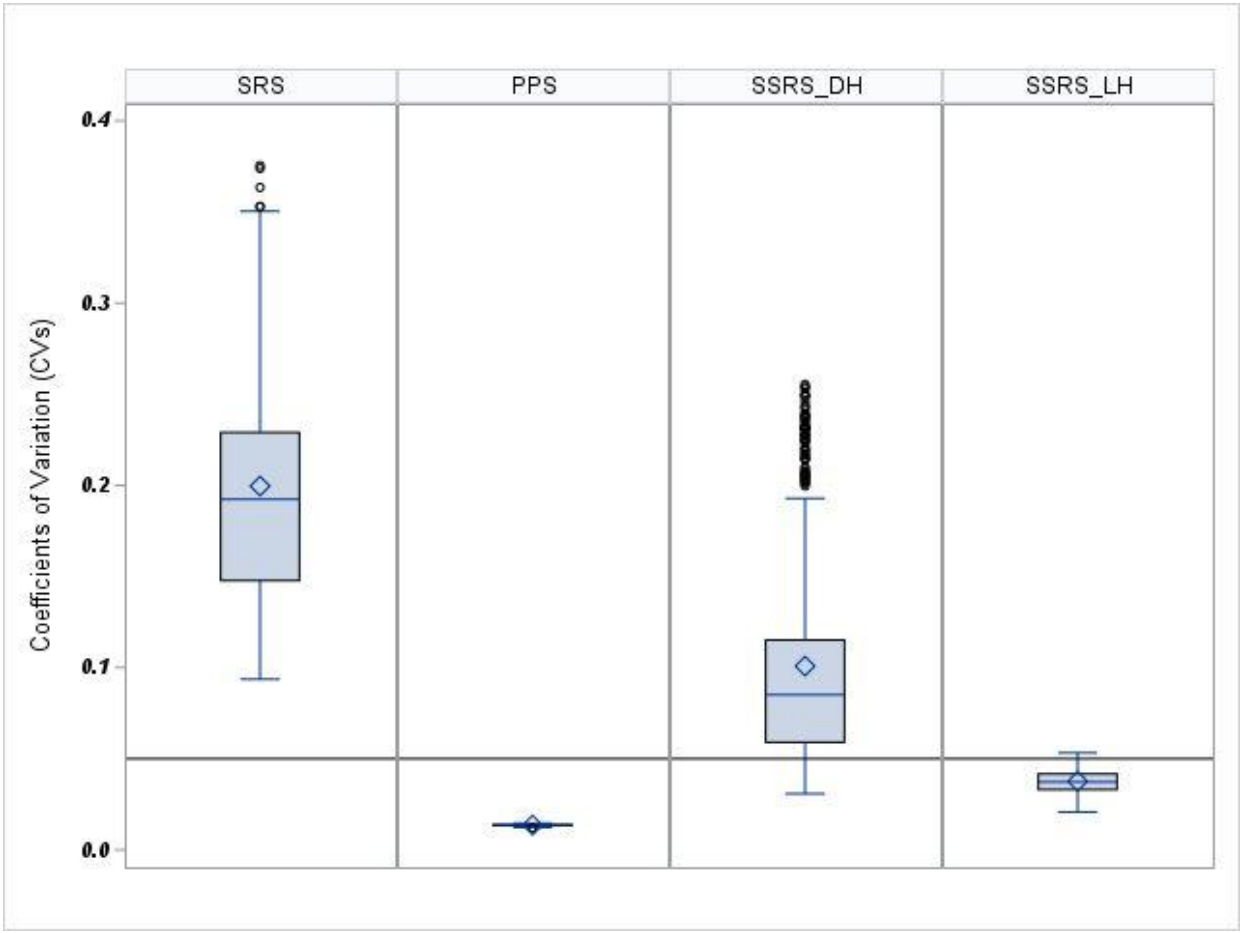
Production Setting Evaluation Approach

- Sample design utilizes frame measure of size **(Annual Payroll)**
  - CANNOT use this variable for evaluation (too good)

- Frame variable **(Total Employment)** used to evaluate effectiveness of design
  - Four candidate sample designs
  - Select 500 samples from the frame using each candidate design
  - Compare coefficients of variation (c.v.'s) using auxiliary frame variable
    - Target c.v. = 0.05 (5%)

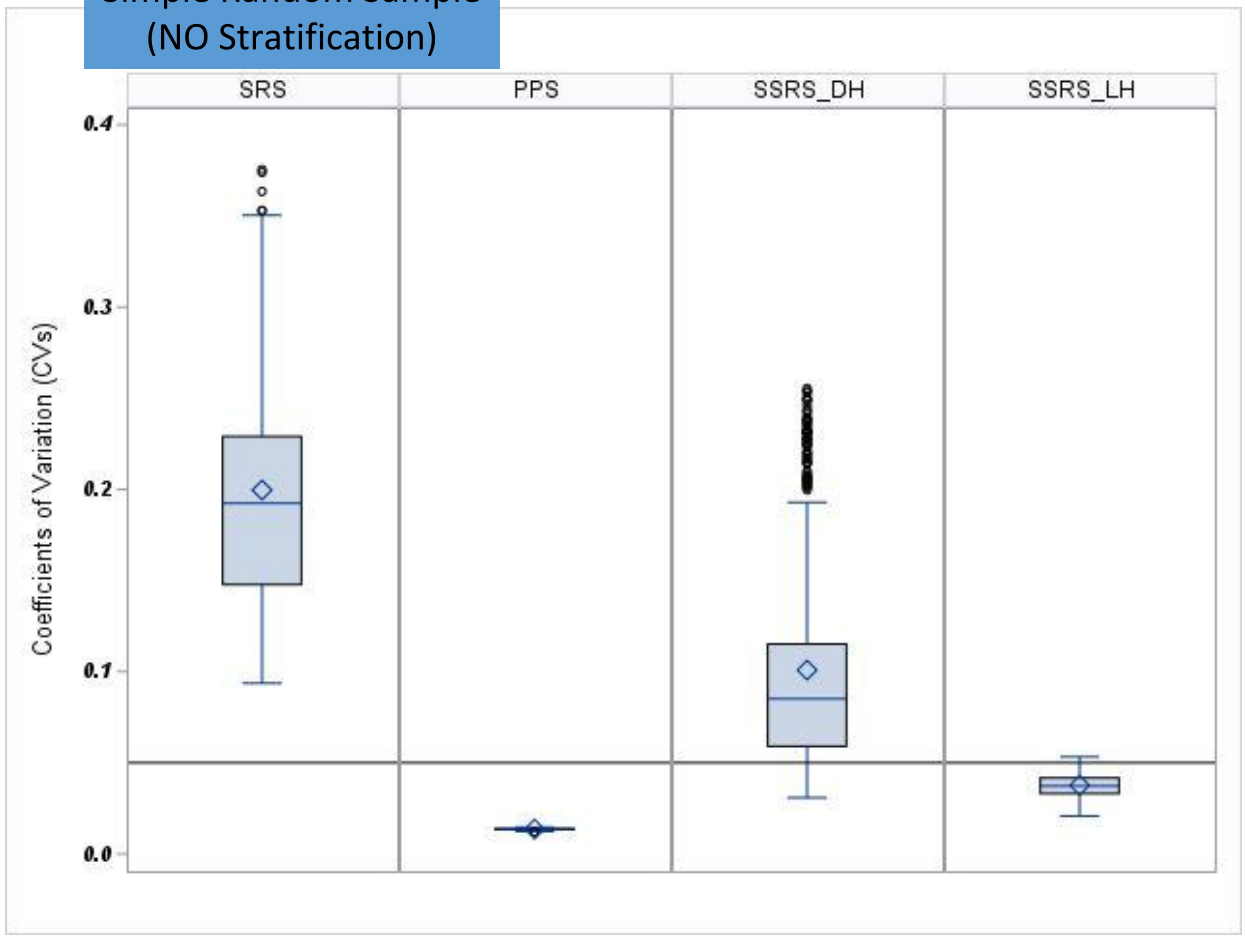
United States® Census Bureau

Results: Total Employment



Results: Total Employment

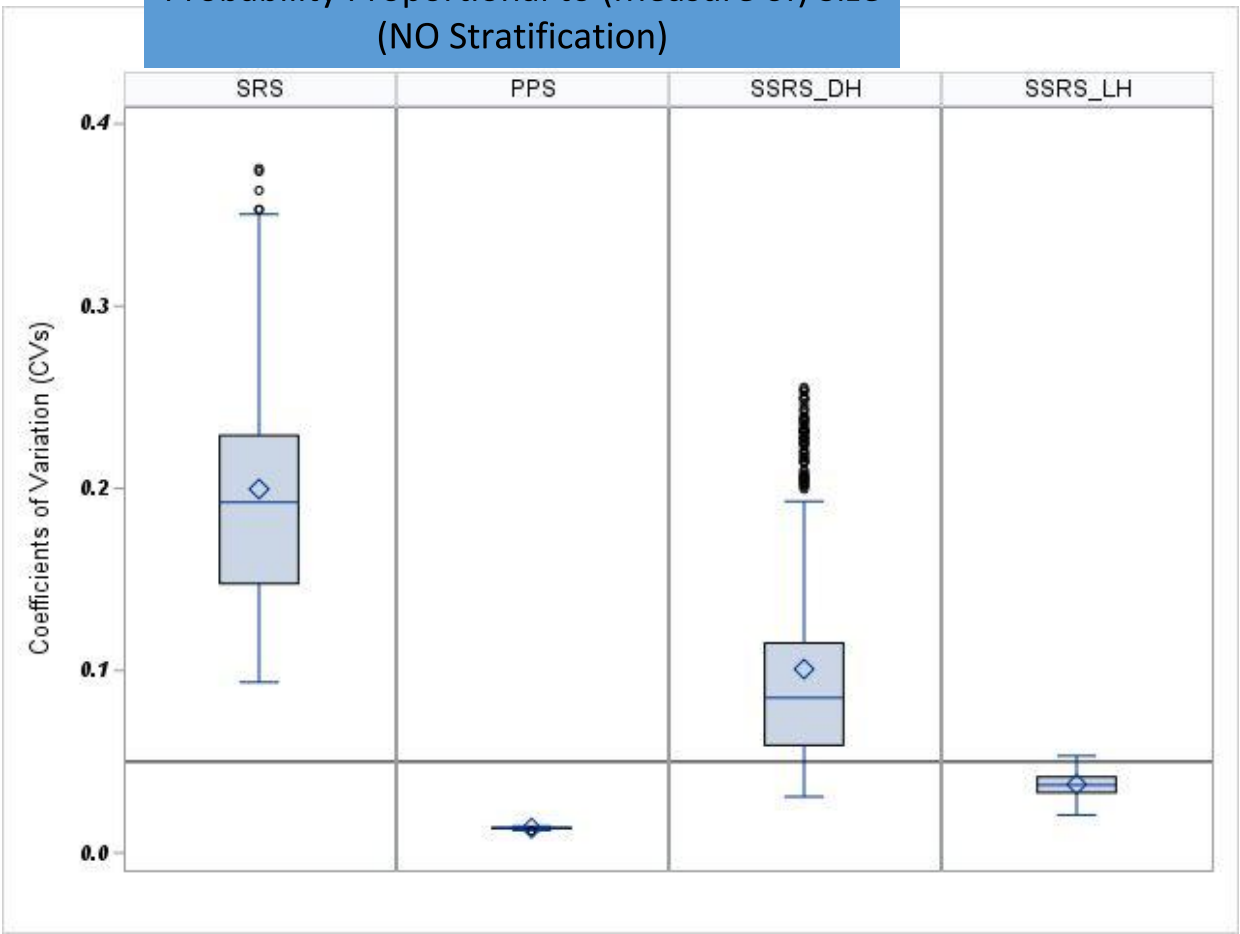
Simple Random Sample
(NO Stratification)



United States® Census Bureau

Results: Total Employment

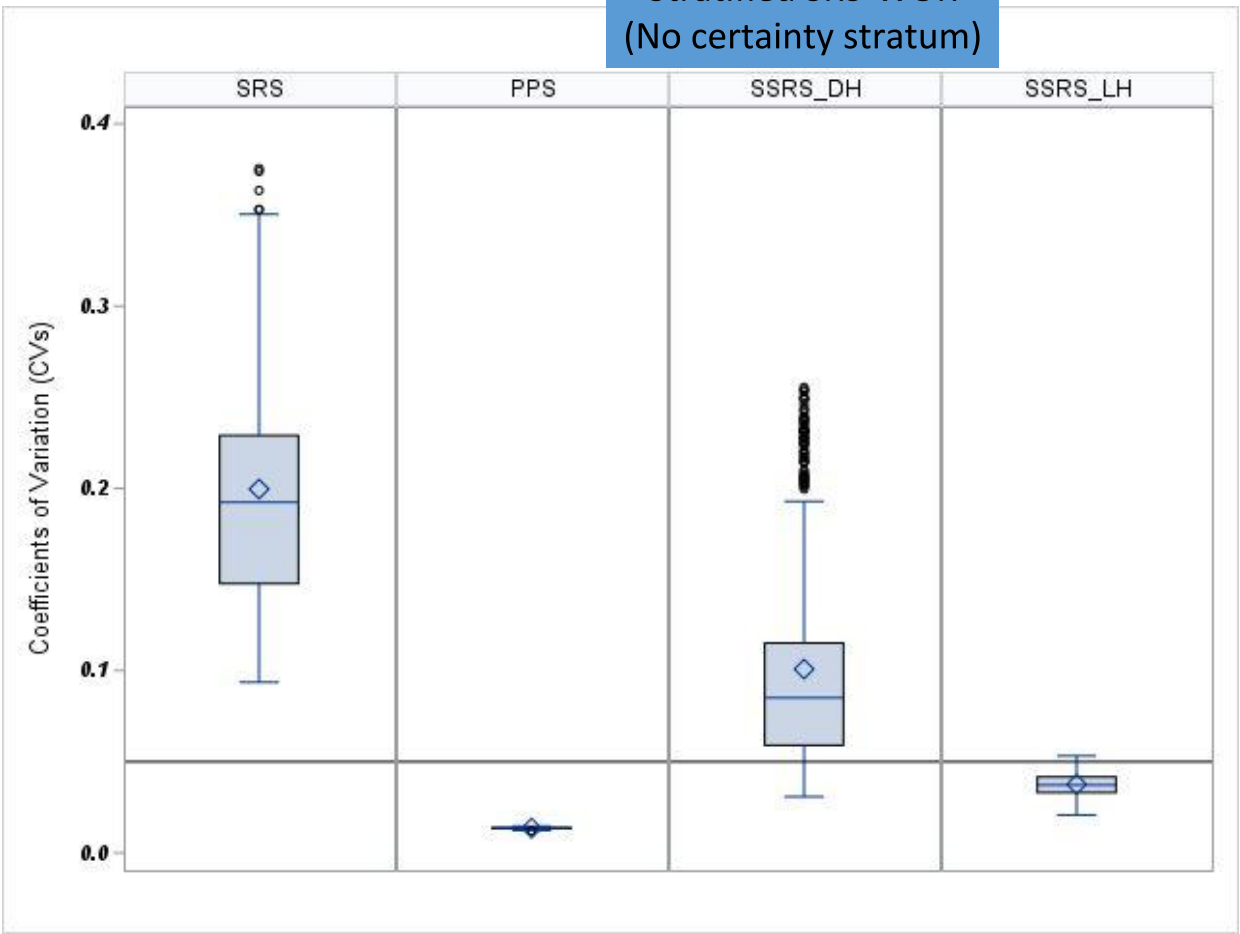
Probability Proportional to (Measure of) Size
(NO Stratification)



United States® Census Bureau

Results: Total Employment

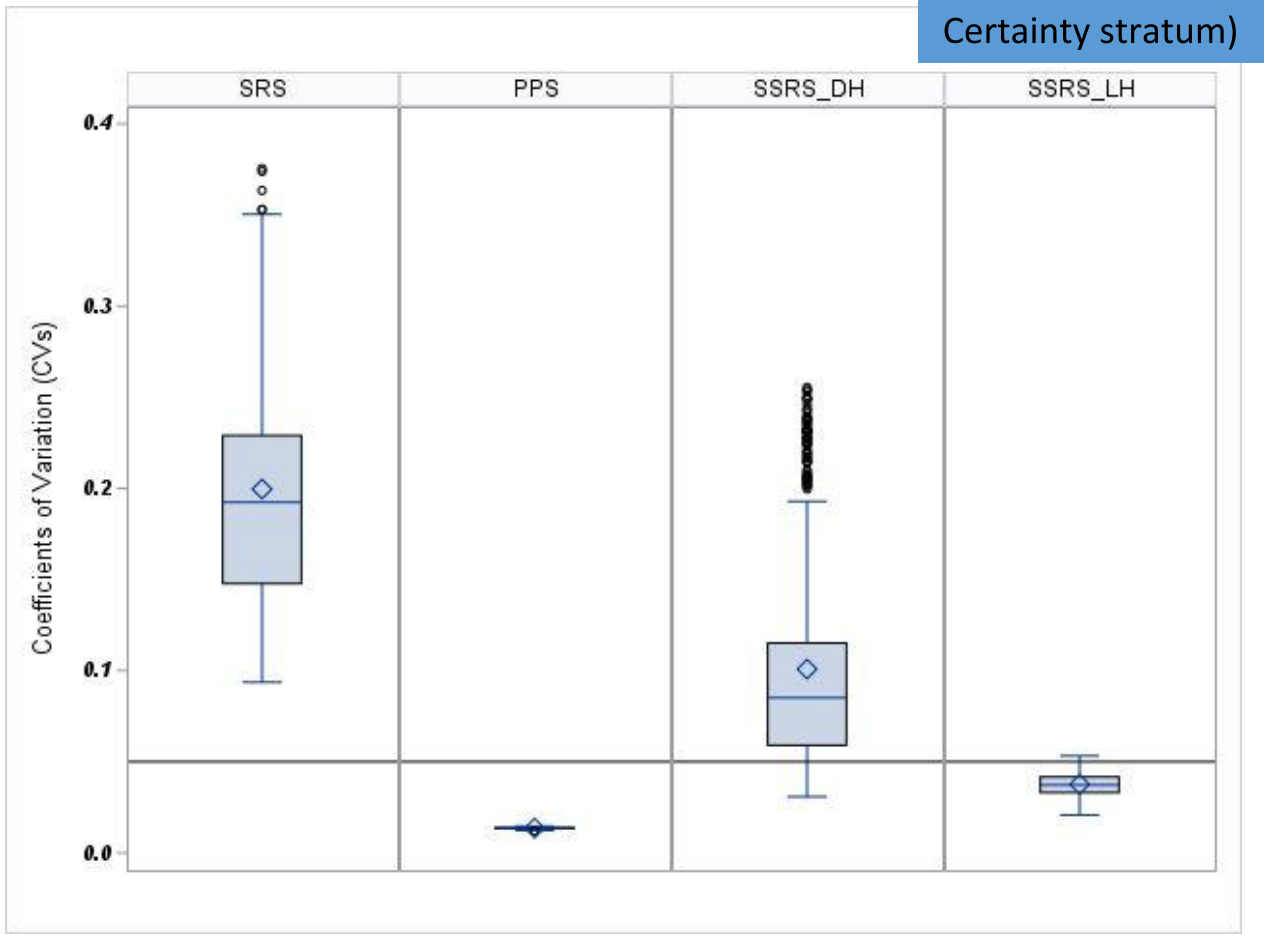
Stratified SRS-WOR
(No certainty stratum)



United States® Census Bureau

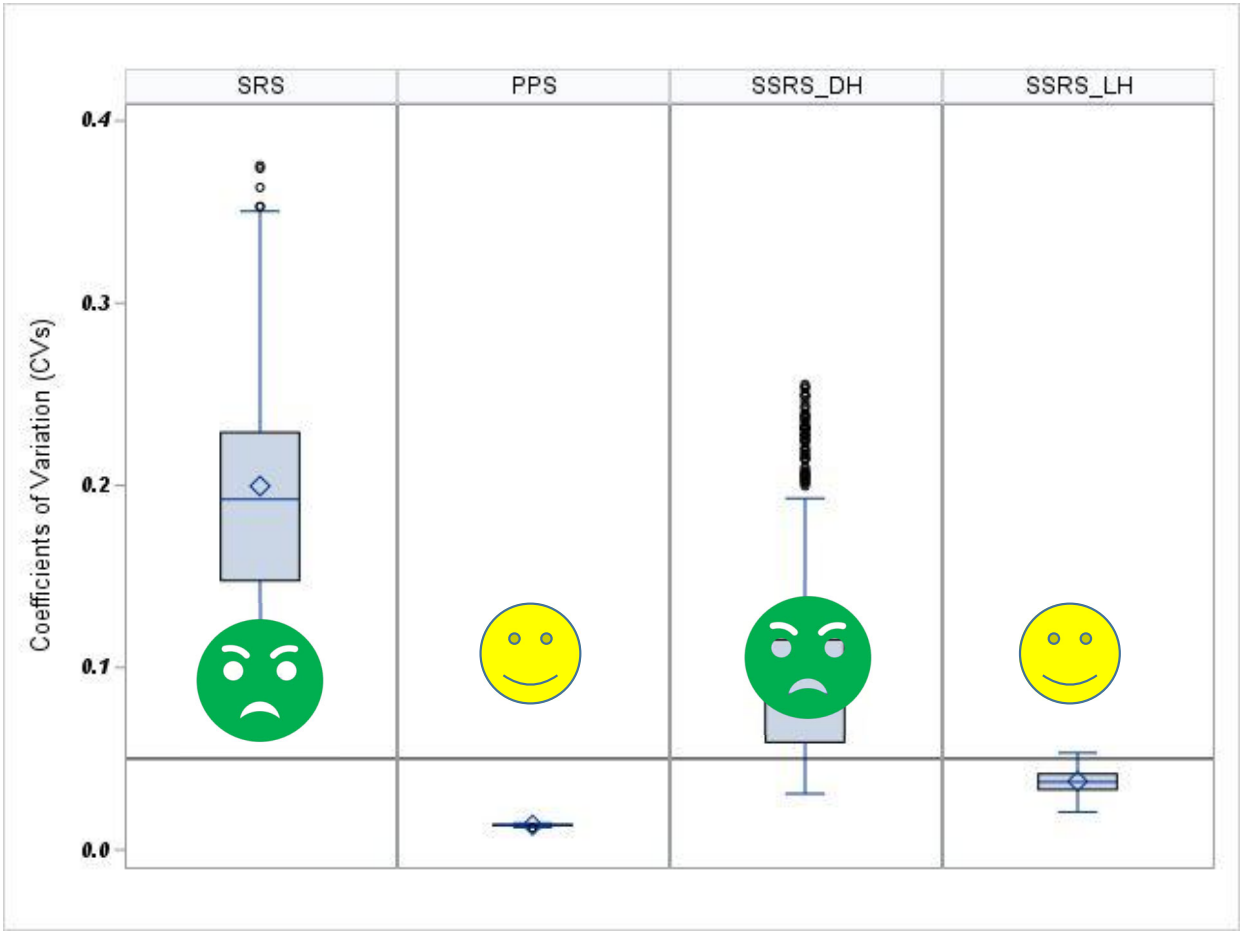
Results: Total Employment

Stratified SRS-WOR
Certainty stratum)

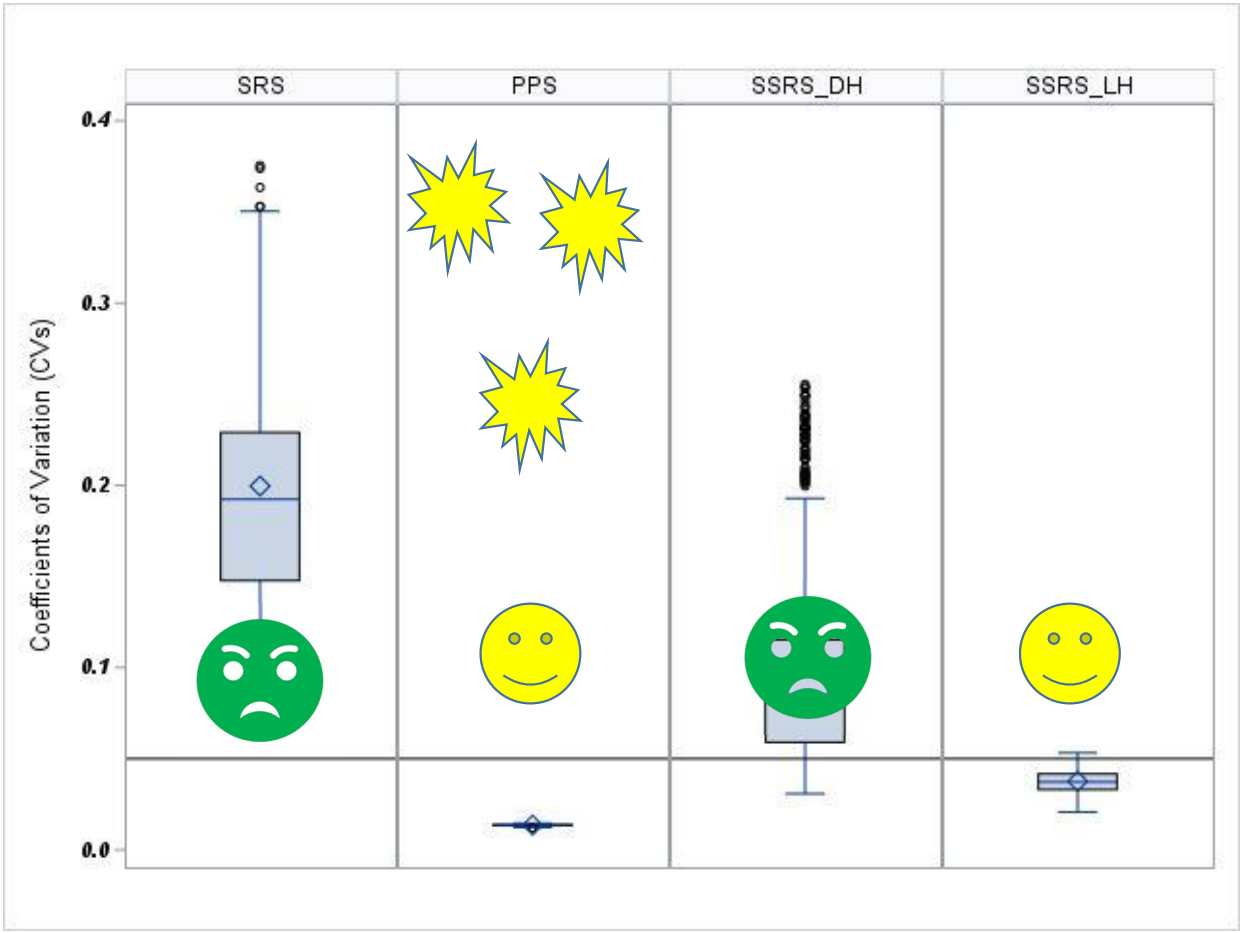


United States® Census Bureau

Results: Total Employment



Results: Total Employment



United States® Census Bureau

But Wait…

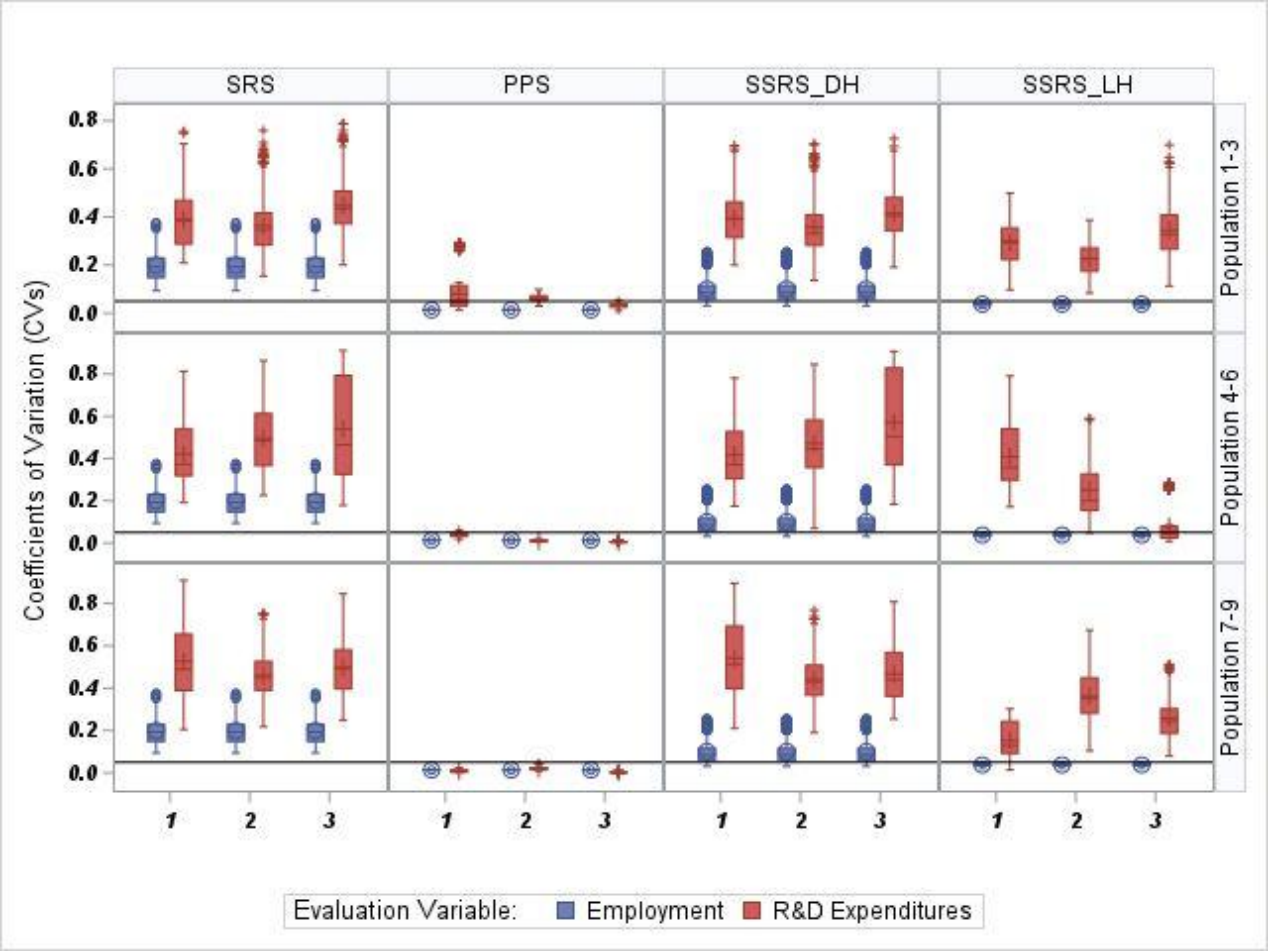
- BERD publishes measures of **Total R&D Expenditures**
  - Weakly related to Annual Payroll (MOS)
  - Weakly related to Total Employment (Auxiliary Frame Variable)

- We have nine synthetic populations
  - Populations 1 through 3: R&D expenditure more likely in smaller units
  - Populations 4 through 6: Weak or no relationship between R&D expenditures and unit size
  - Populations 7 through 9: R&D expenditure more likely in larger units

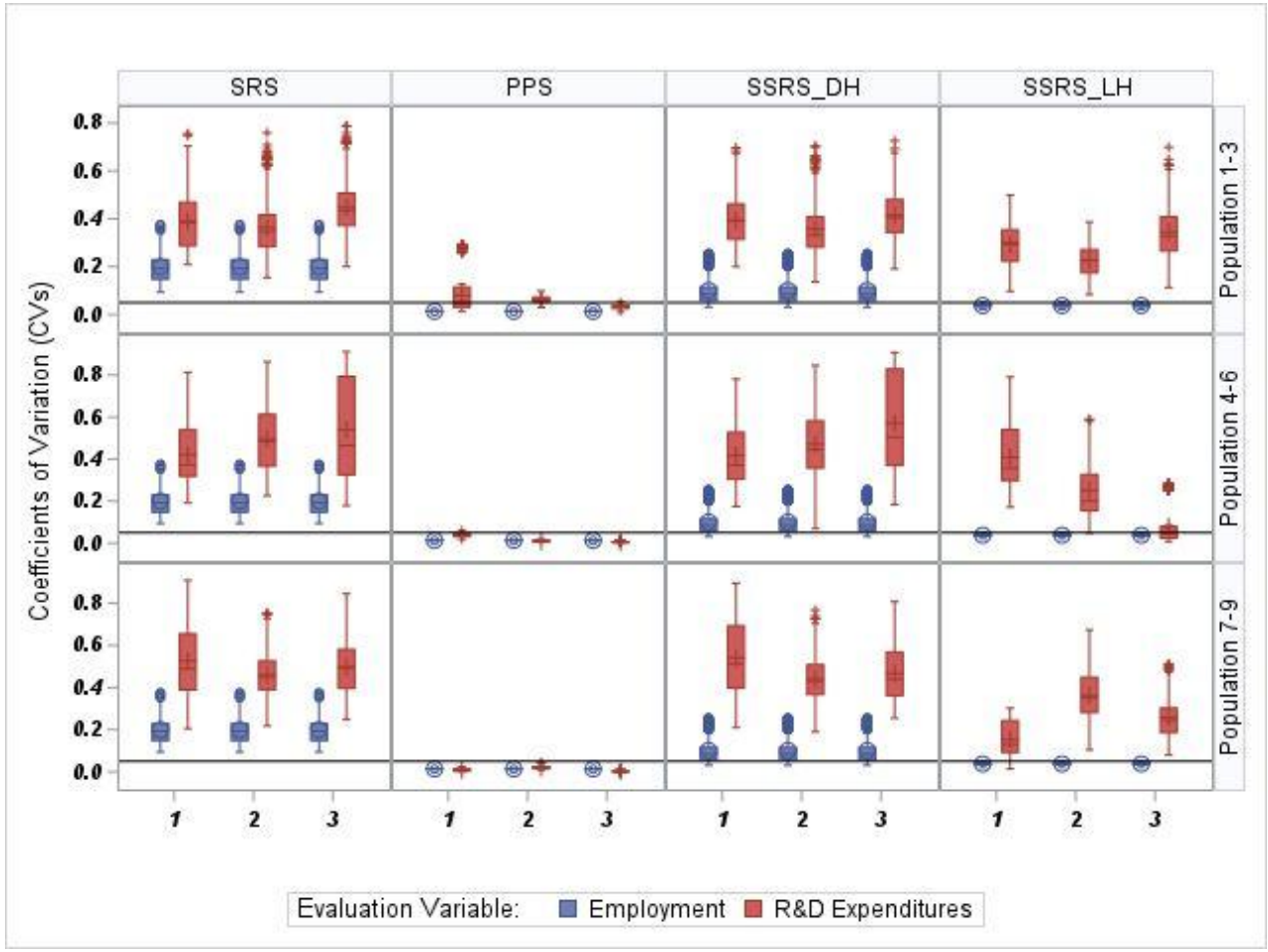
- Let's compare…

United States® Census Bureau

CV(Total Employment) Vs CV(Total R&D Expenditures)



CV(Total Employment) Vs CV(Total R&D Expenditures)



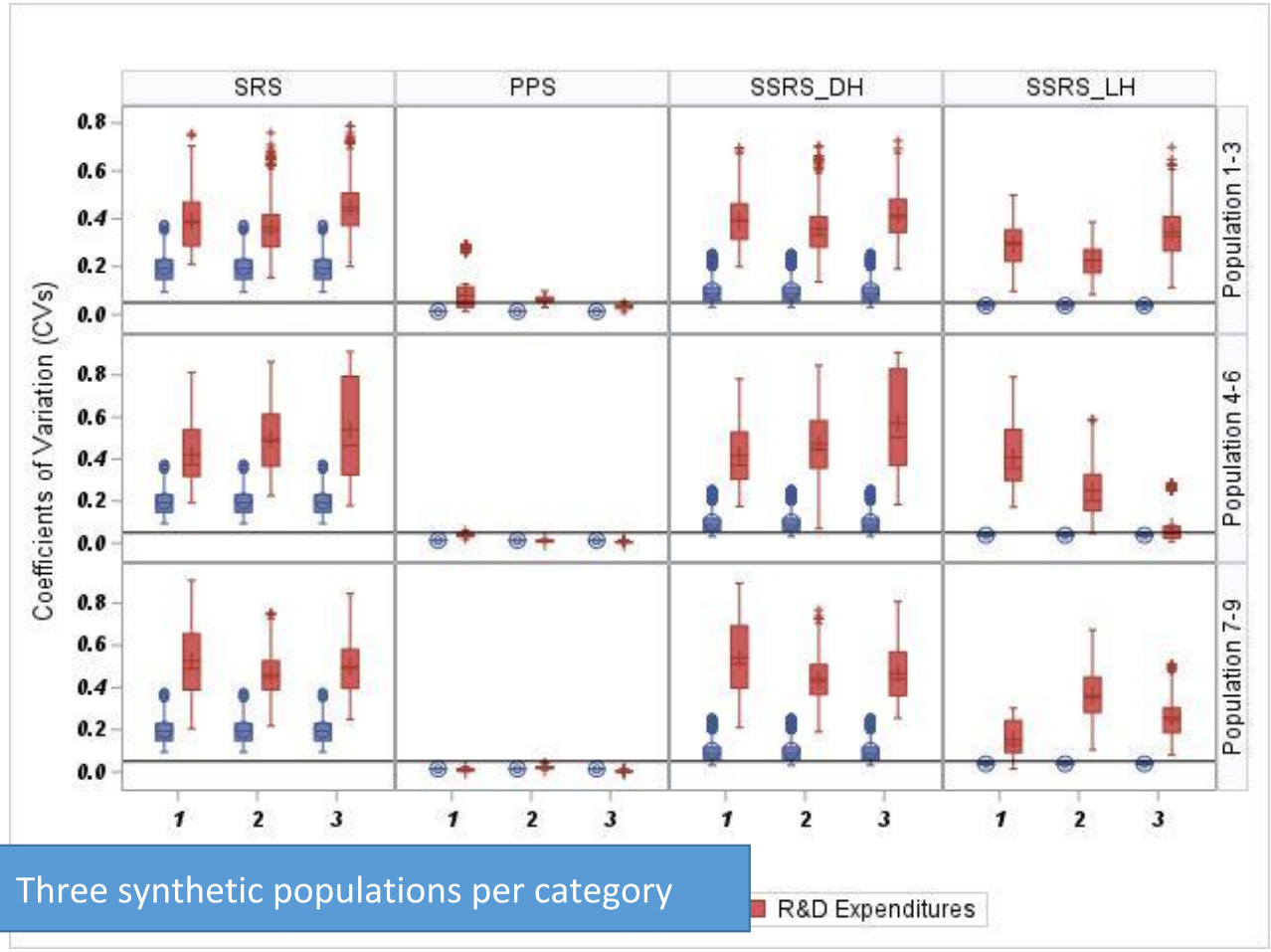
R&D in smaller units

R&D throughout population

R&D in larger units

United States® Census Bureau

CV(Total Employment) Vs CV(Total R&D Expenditures)



R&D in smaller units

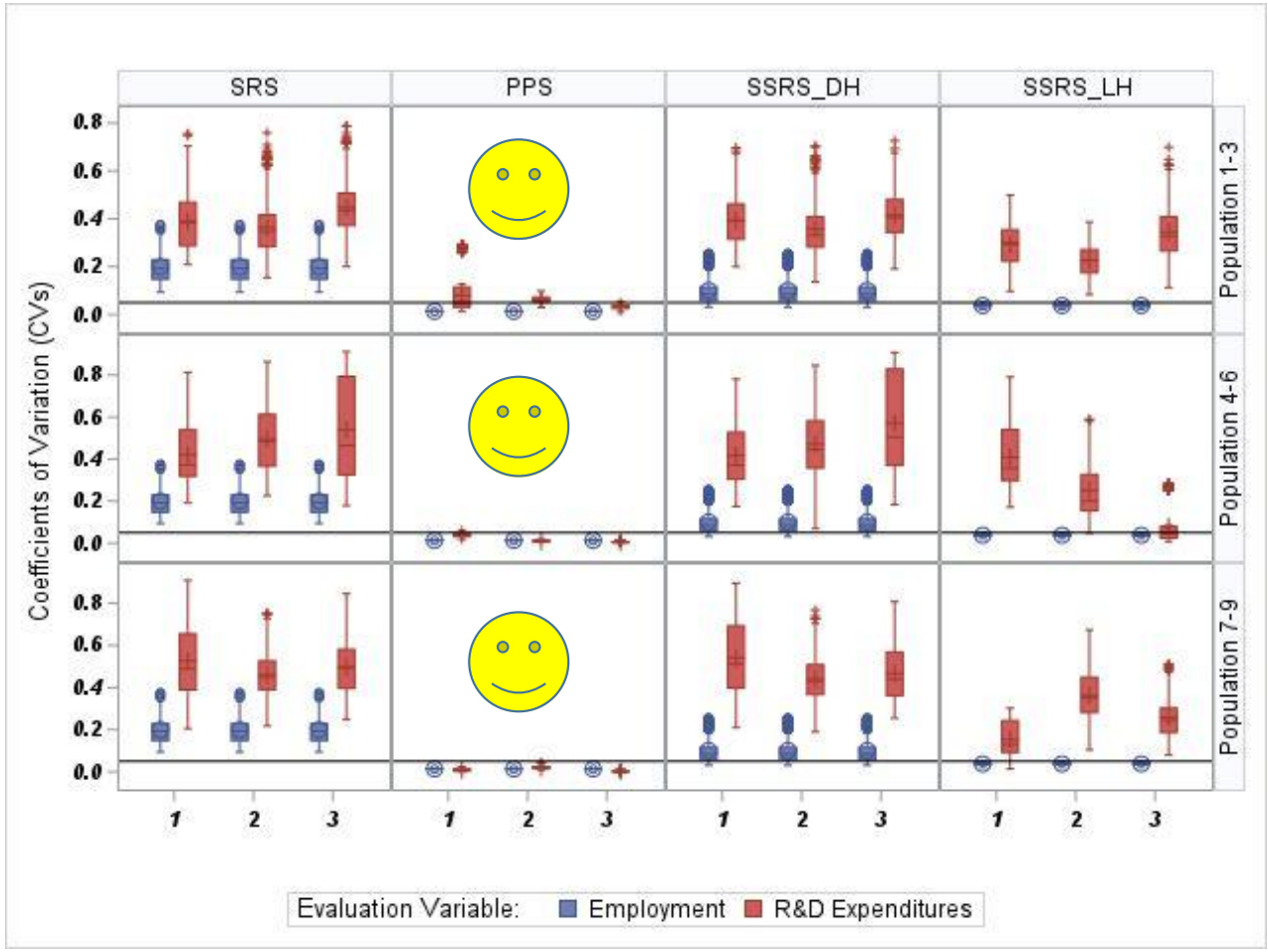
R&D throughout population

R&D in larger units

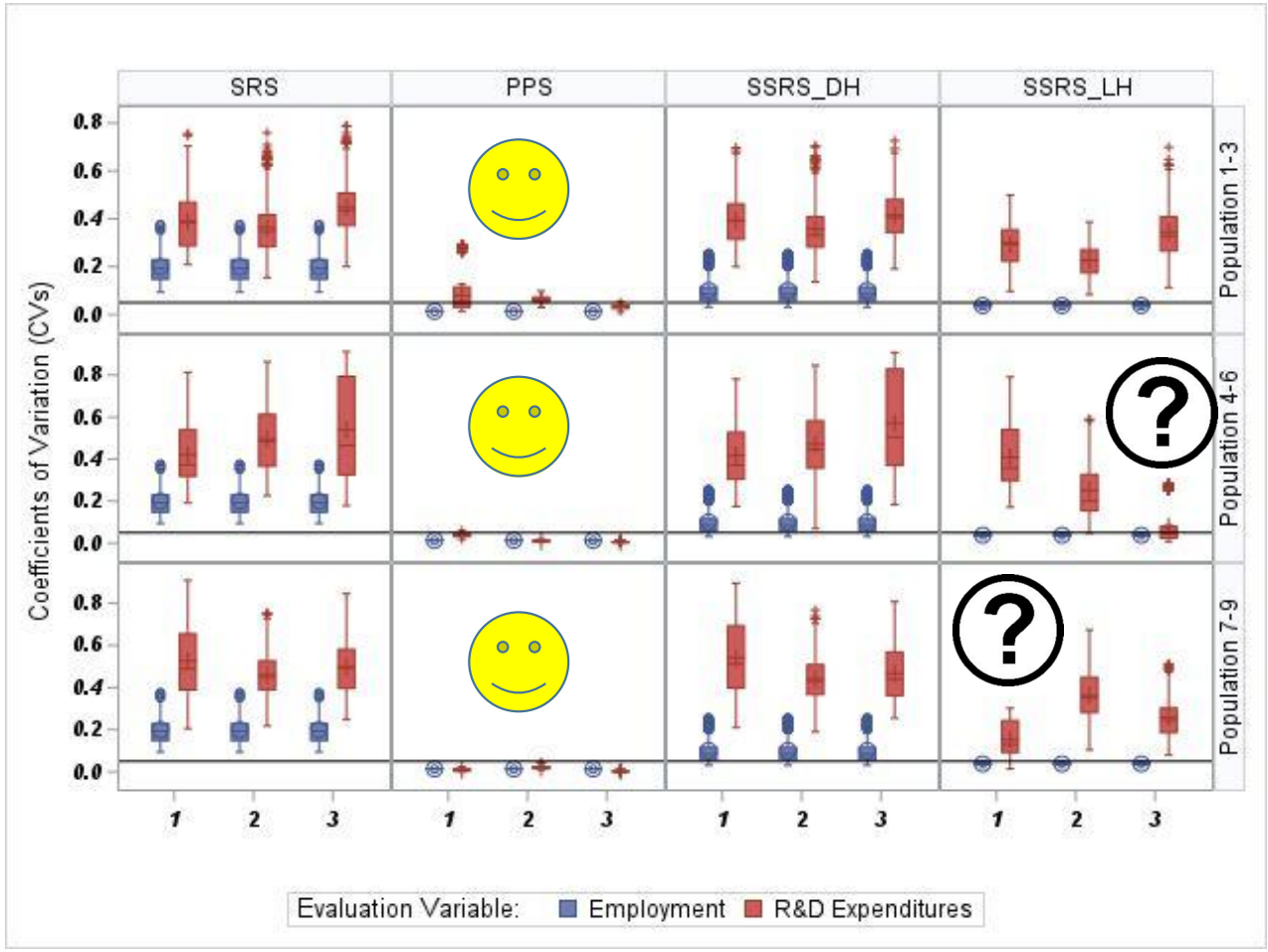
Three synthetic populations per category

United States® Census Bureau

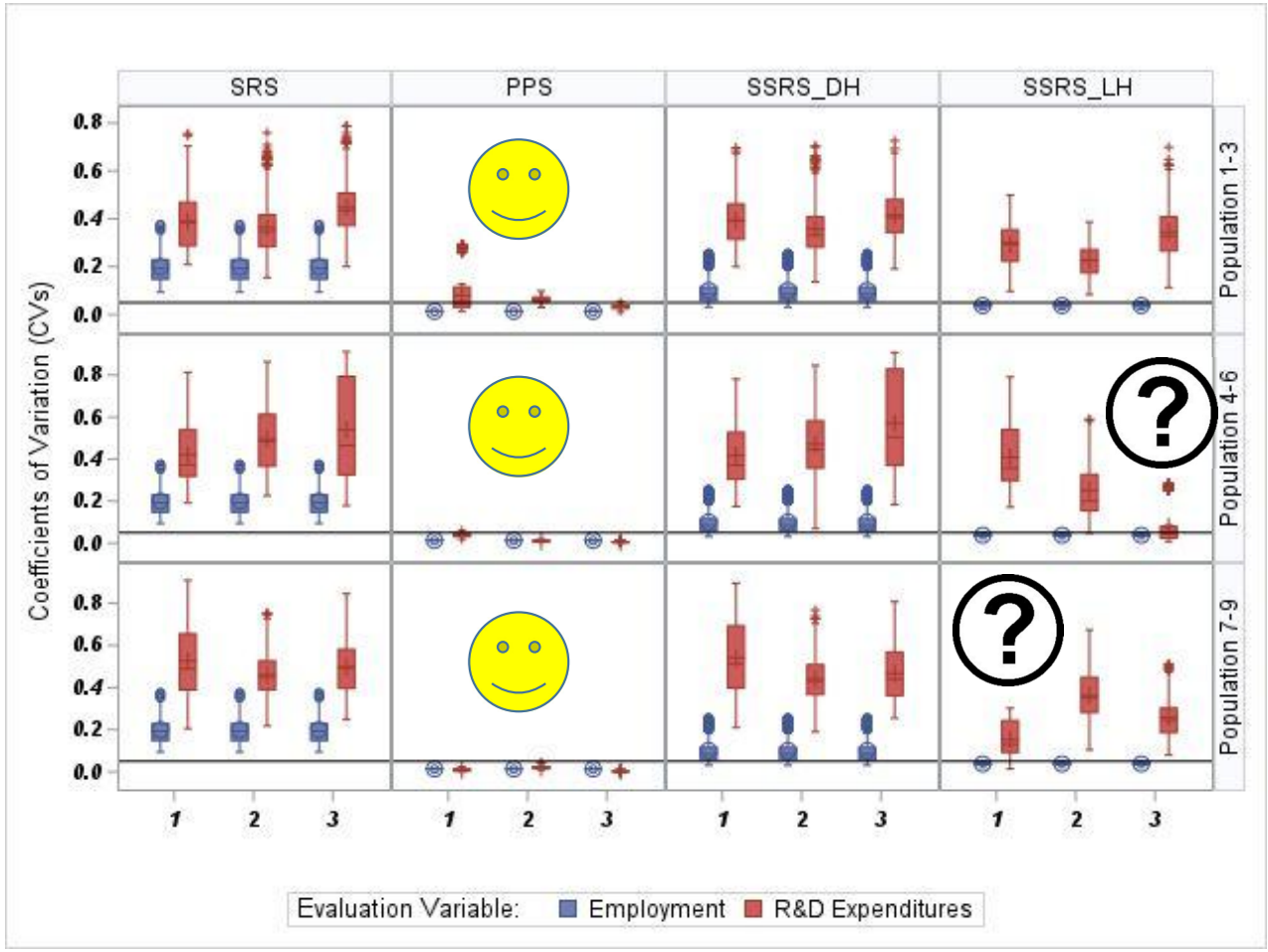
CV(Total Employment) Vs CV(Total R&D Expenditures)



CV(Total Employment) Vs CV(Total R&D Expenditures)



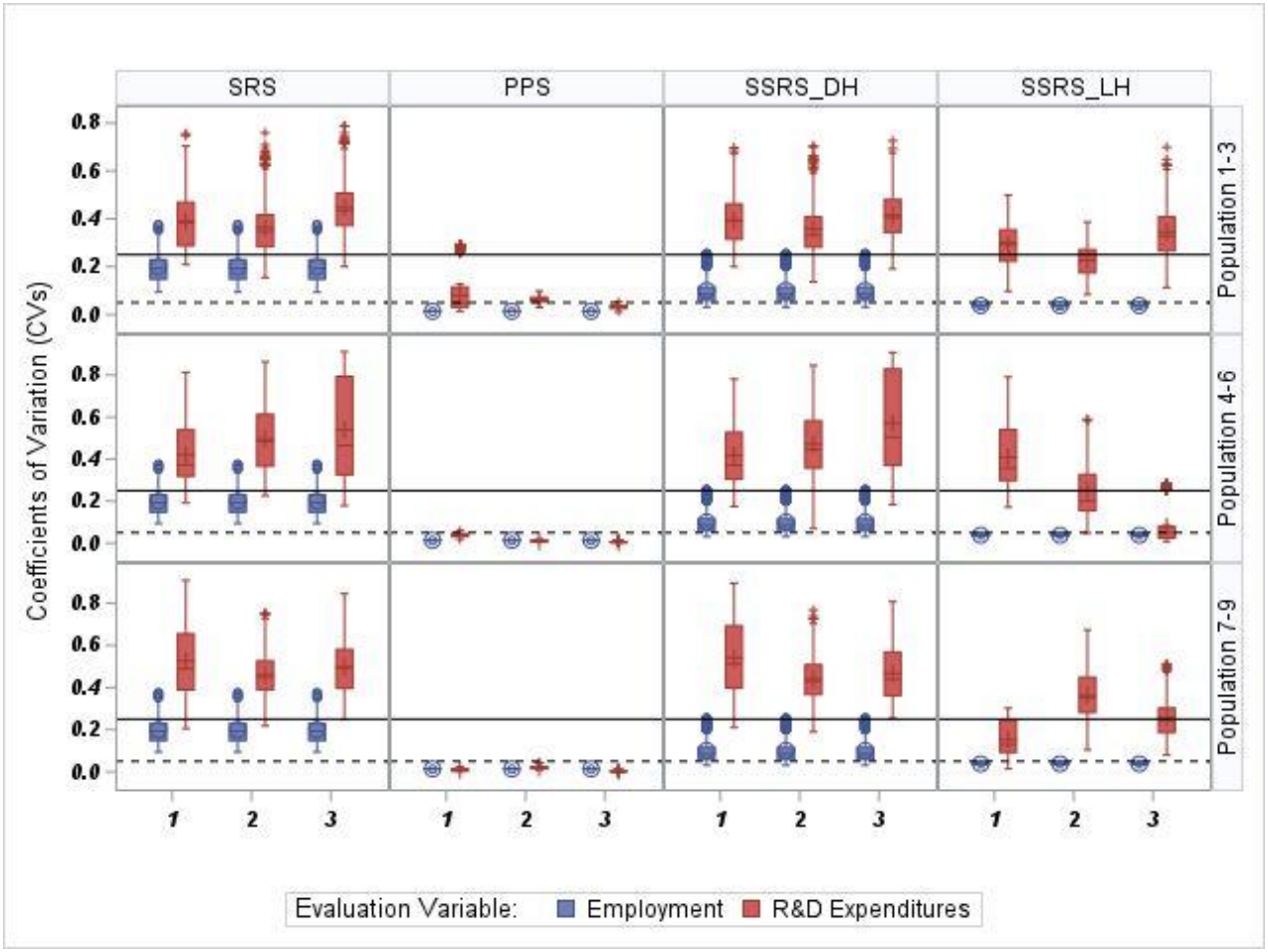
CV(Total Employment) Vs CV(Total R&D Expenditures)



Raise C.V. target since we are assessing the design for a rare characteristic

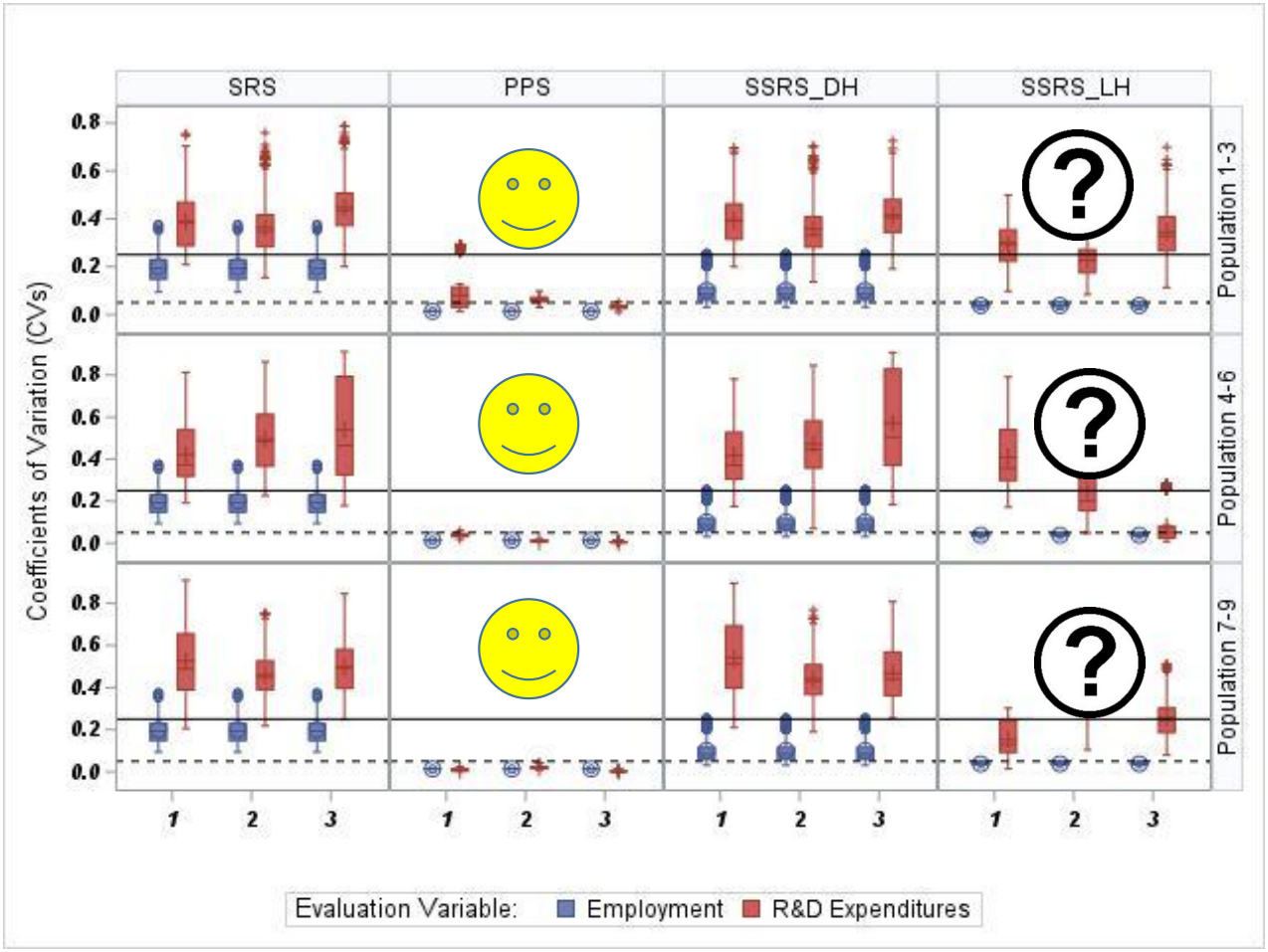
United States® Census Bureau

CV(Total Employment) Vs CV(Total R&D Expenditures)



United States® Census Bureau

CV(Total Employment) Vs CV(Total R&D Expenditures)



United States® Census Bureau

Done?

- Clearly, the PPS design is “best”
  - Consistently lowest c.v.’s
  - Little variability between samples (by design)

- But…we have nine synthetic populations

- We can look at statistical properties over repeated samples
  - Assessed precision
  - Assess accuracy (bias)
  - Fitness of data for inference (90% confidence interval coverage)

United States® Census Bureau

Evaluation Statistics

- Relative Bias of the Estimate

$$RB\left(\hat{\theta}_p^d\right) = \frac{\sum_{s=1}^{500} \hat{\theta}_{p(s)}^d}{\theta_p}$$

- 90% Confidence Interval Coverage Rate

$$\mathrm{CR}\left(\hat{\theta}_p^d\right) = \text{percentage of 90\% CI's with sample design } d \text{ in population } p \text{ that contain } \theta_p$$

United States® Census Bureau

Evaluation Statistics

- Relative Bias of the Estima

$$RB\left(\hat{\theta}_p^d\right) = \frac{\sum_{s=1}^{500} \hat{\theta}_{p(s)}^d}{\theta_p}$$

True value of statistic in synthetic population *p*

- Prevalence (proportion) of companies that perform R&D activities
- Total R&D expenditures

- 90% Confidence Interval Coverage Rate

$$\mathrm{CR}\left(\hat{\theta}_p^d\right) = \text{percentage of 90\% CI's with sample design } d \text{ in population } p \text{ that contain } \theta_p$$

United States® Census Bureau

Evaluation Statistics

- Relative Bias of the Estima

$$RB\left(\hat{\theta}_p^d\right) = \frac{\sum_{s=}^{50(} \hat{\theta}_{p(s)}^d}{\theta_p}$$

Sample design *d* (*d = SRS, PPS, SSRS_DH, SSRS_LH*)
p = partially synthetic frame (*p*= 1, …, 9)
s = independent sample (*s* = 1, …, 500)

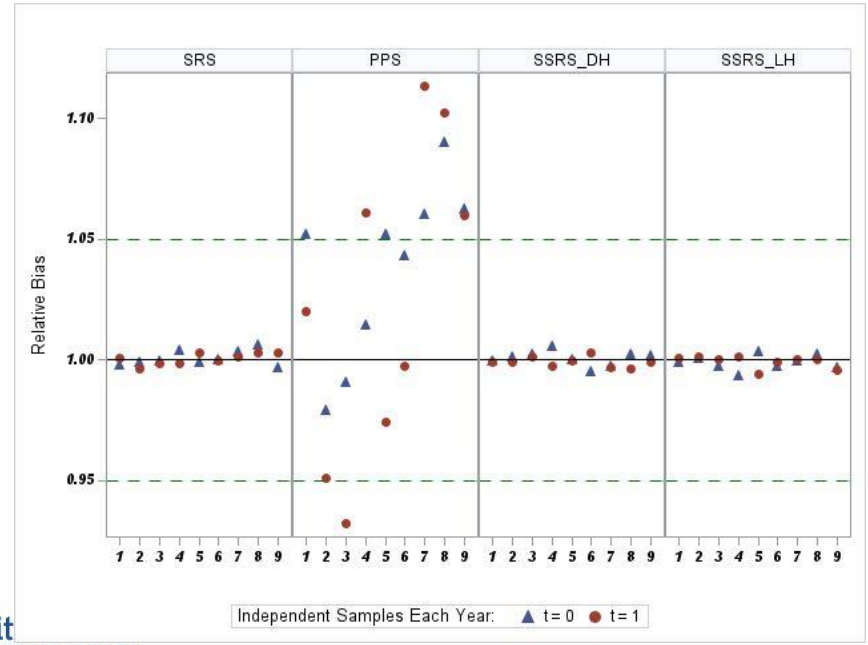
- 90% Confidence Interval Coverage Rate

$$\mathrm{CR}\left(\hat{\theta}_p^d\right) = \text{percentage of 90\% CI's with sample design } d \text{ in population } p \text{ that contain } \theta_p$$

United States® Census Bureau

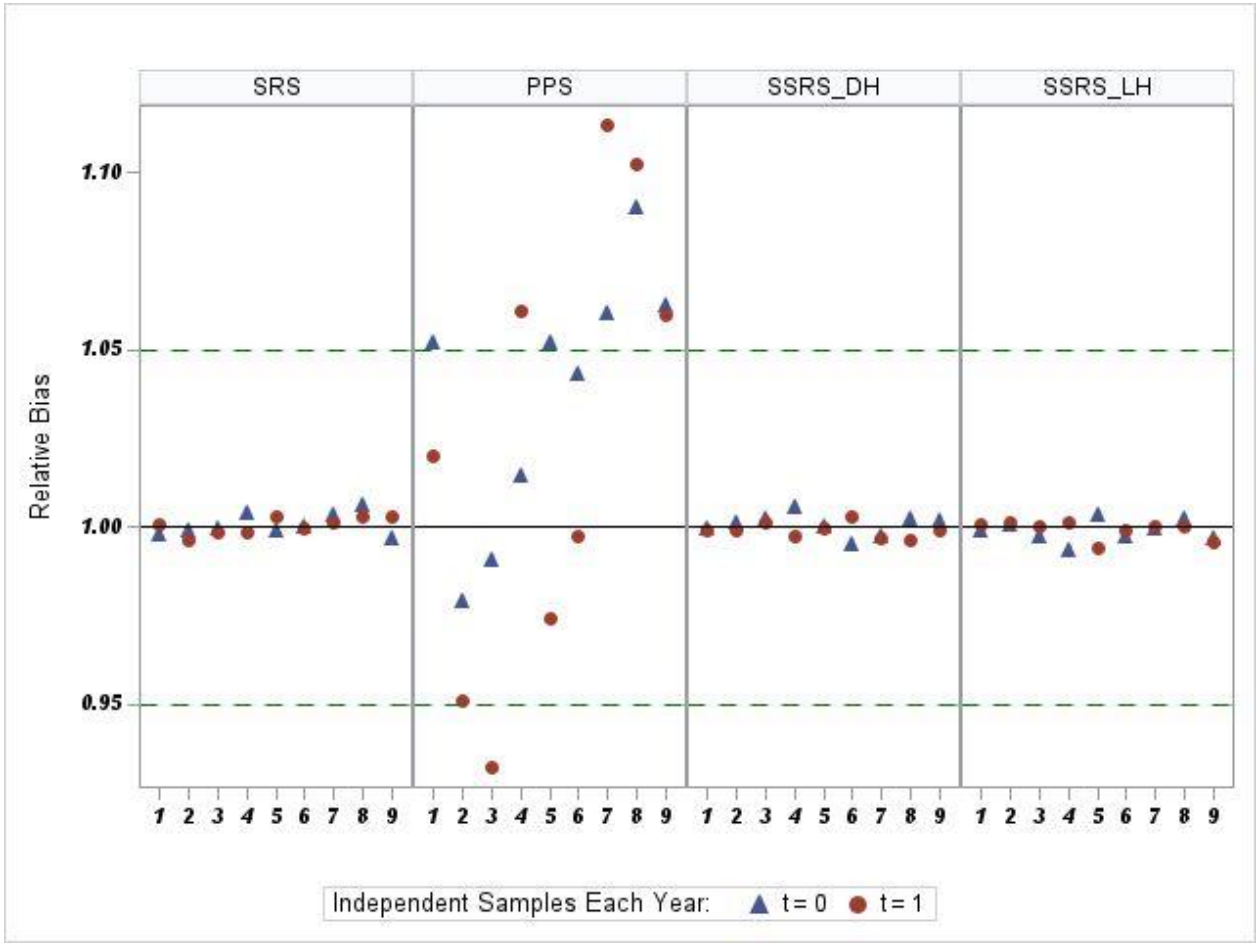
Relative Bias: "New Survey" (Case 1)

R&D Prevalence (Proportion with R&D)

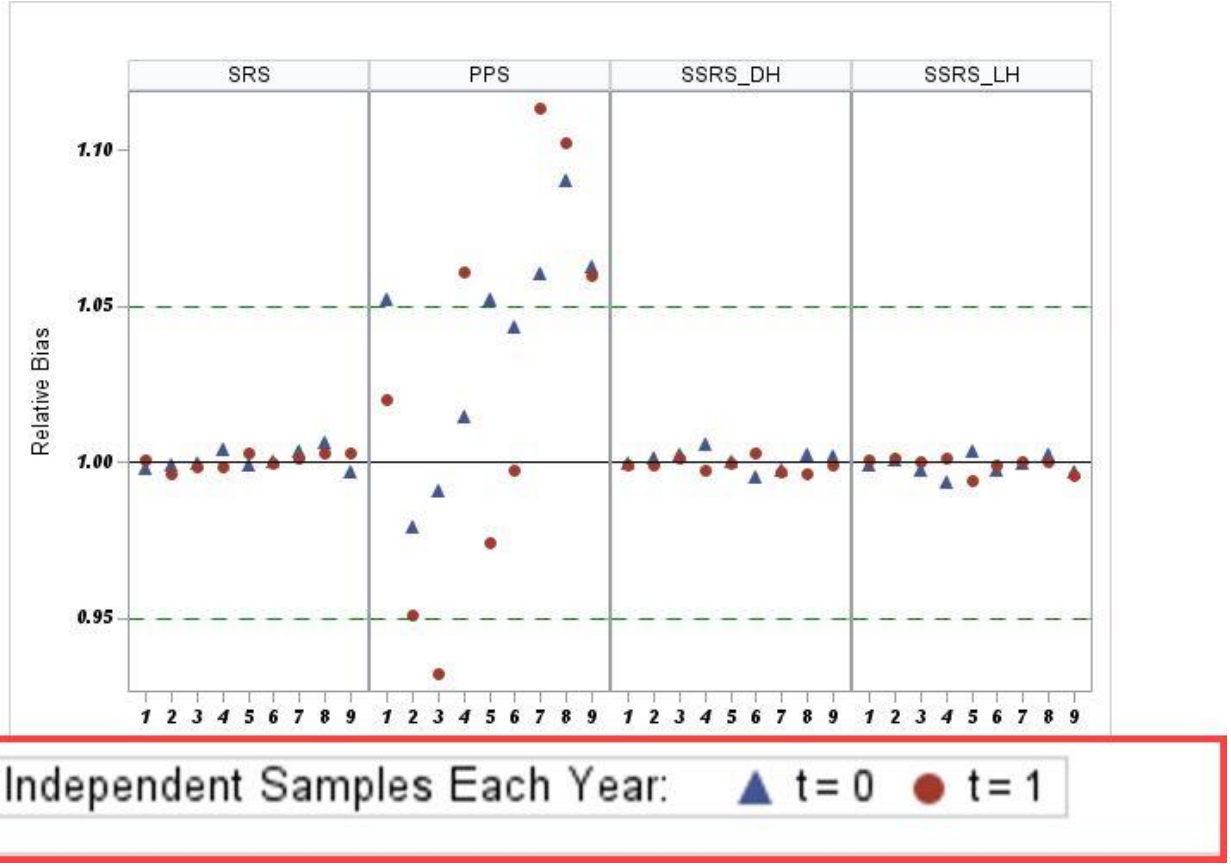


Unit Census Bureau

Relative Bias: "New Survey" (Case 1)

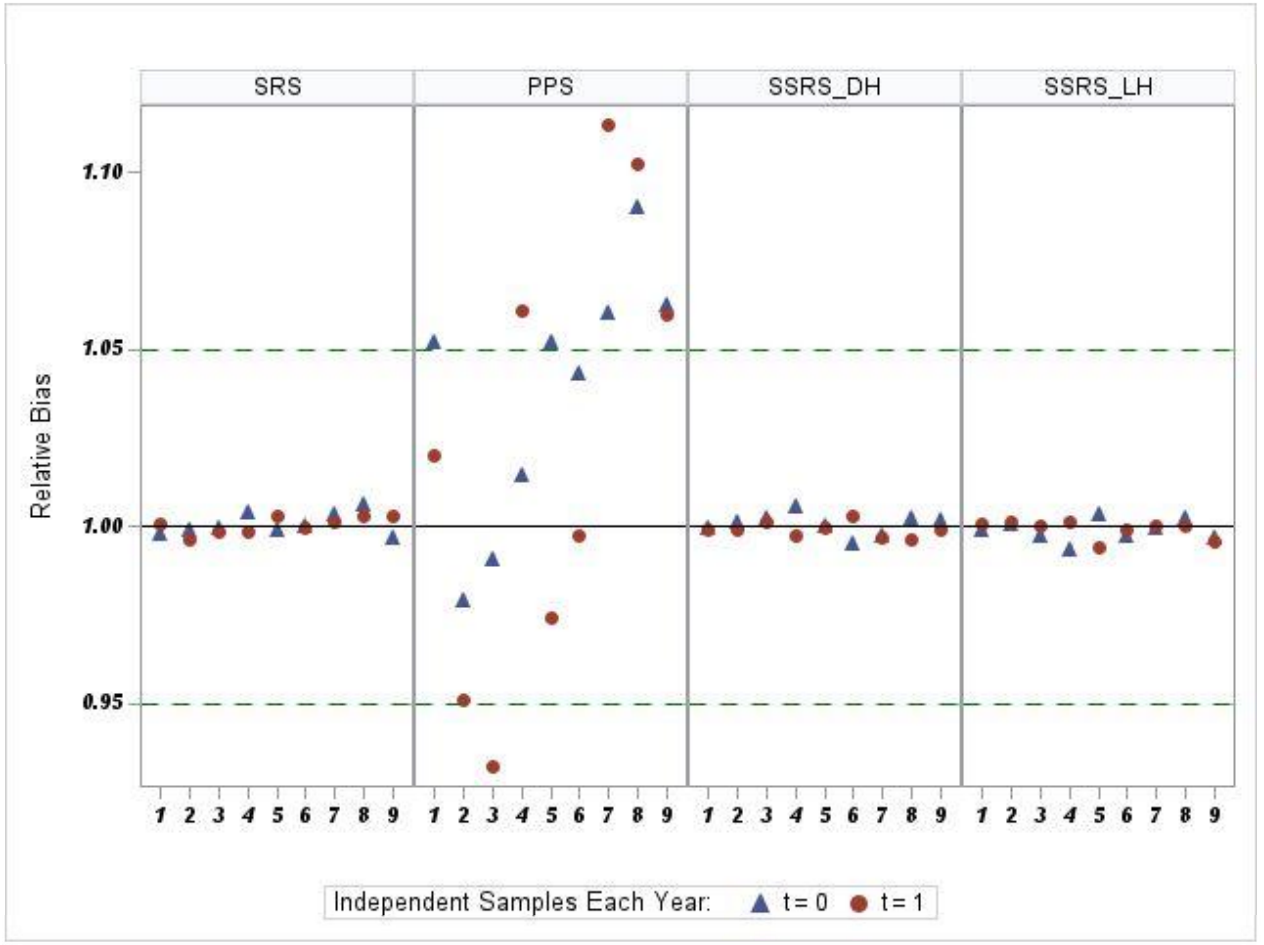


Relative Bias: "New Survey" (Case 1)



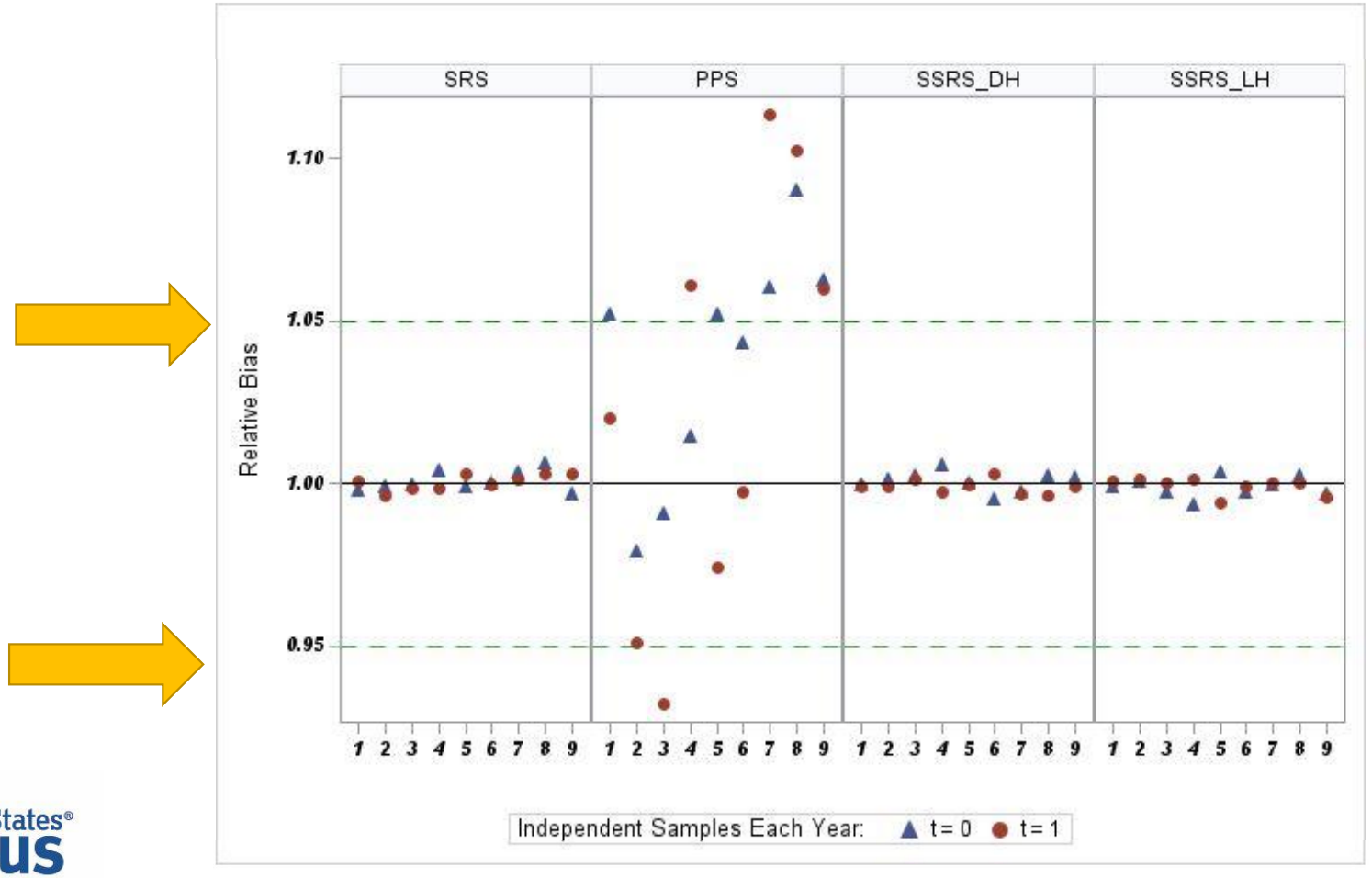
Relative Bias: "New Survey" (Case 1)

Unbiased



R&D Prevalence (Proportion with R&D)

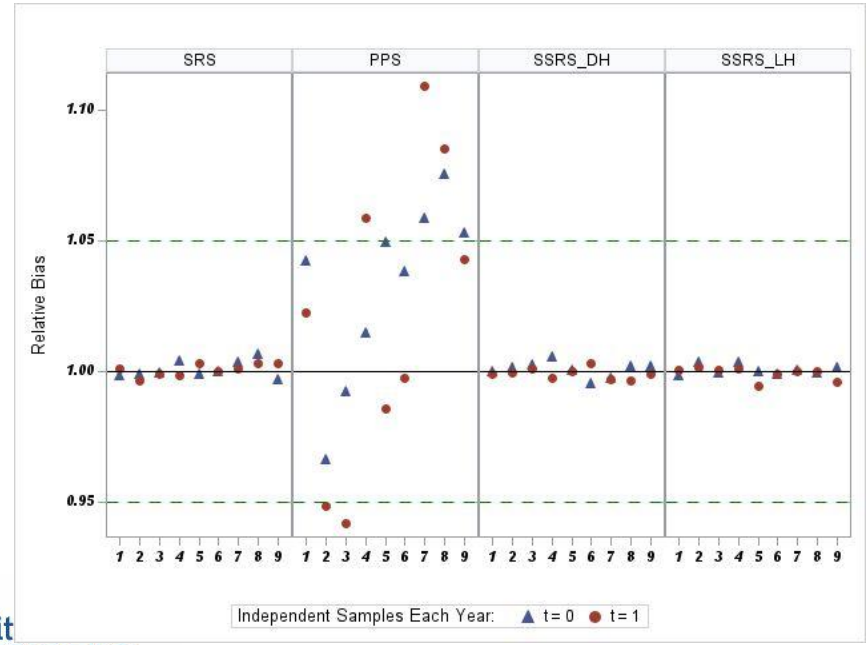
Relative Bias: "New Survey" (Case 1)



R&D Prevalence (Proportion with R&D)

Relative Bias: "New Survey" (Case 1)

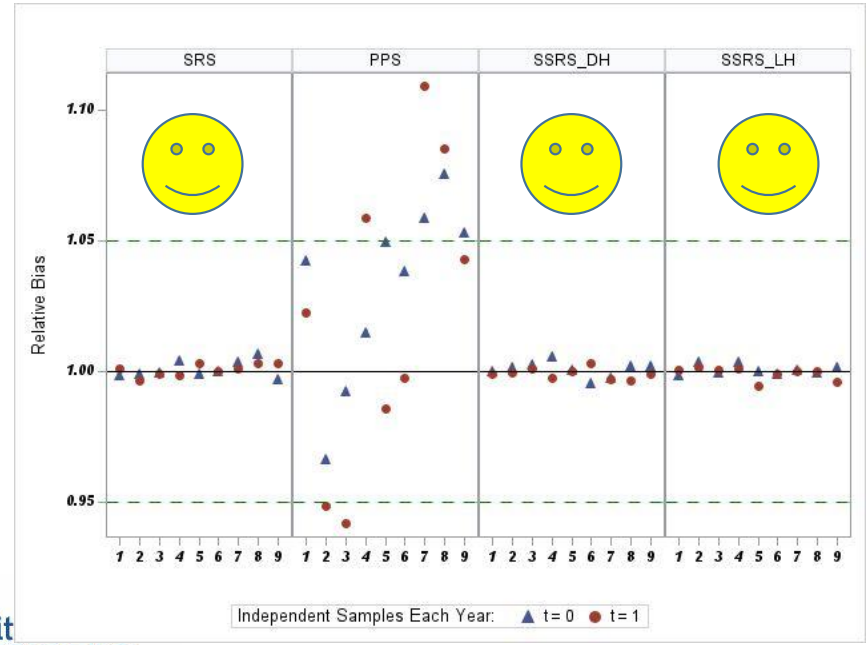
R&D Prevalence (Proportion with R&D)



Unit Census Bureau

Relative Bias: "New Survey" (Case 1)

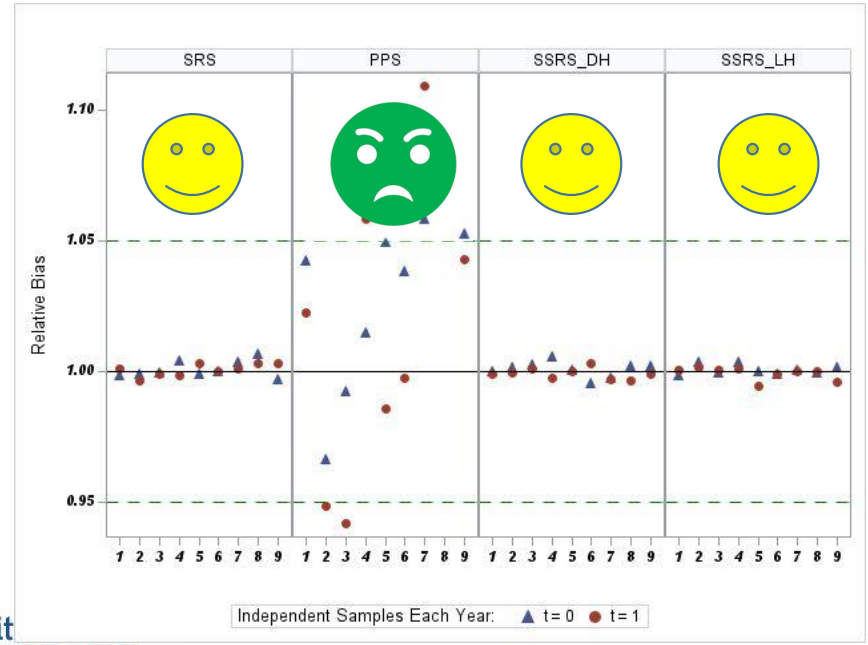
R&D Prevalence (Proportion with R&D)



Unit Census Bureau

Relative Bias: "New Survey" (Case 1)

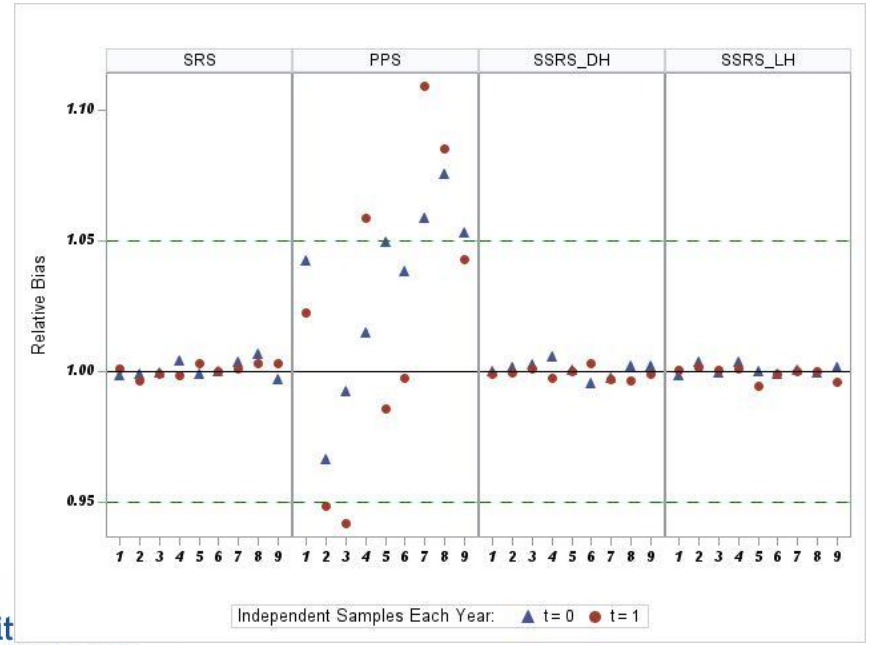
R&D Prevalence (Proportion with R&D)



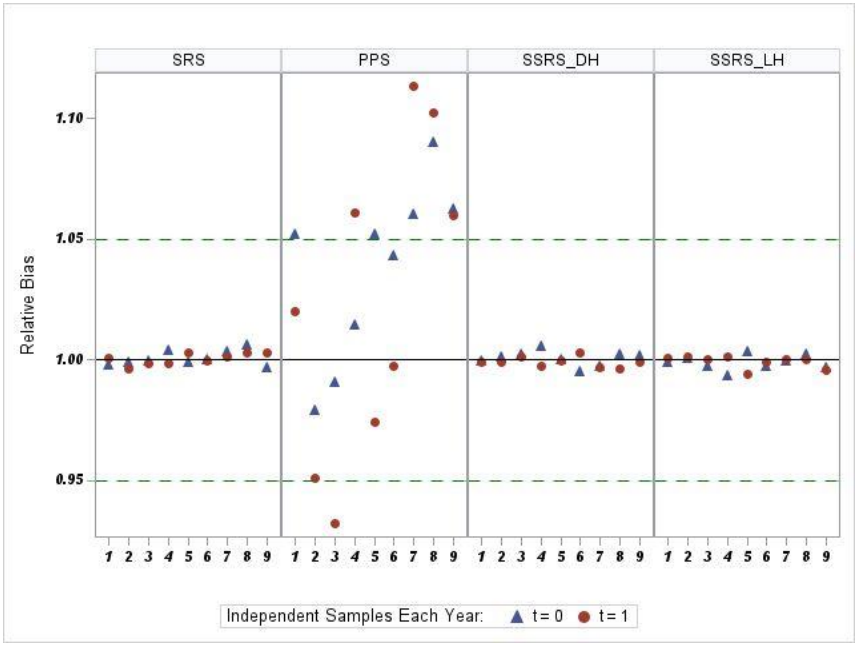
Unit Census Bureau

Relative Bias: "New Survey" (Case 1)

R&D Prevalence (Proportion with R&D)



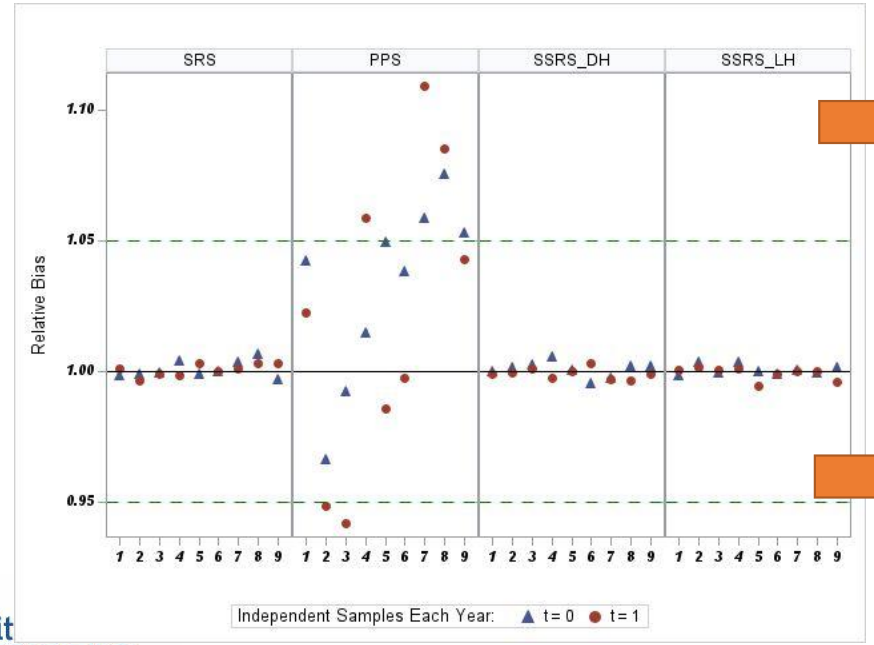
Total R&D Expenditures



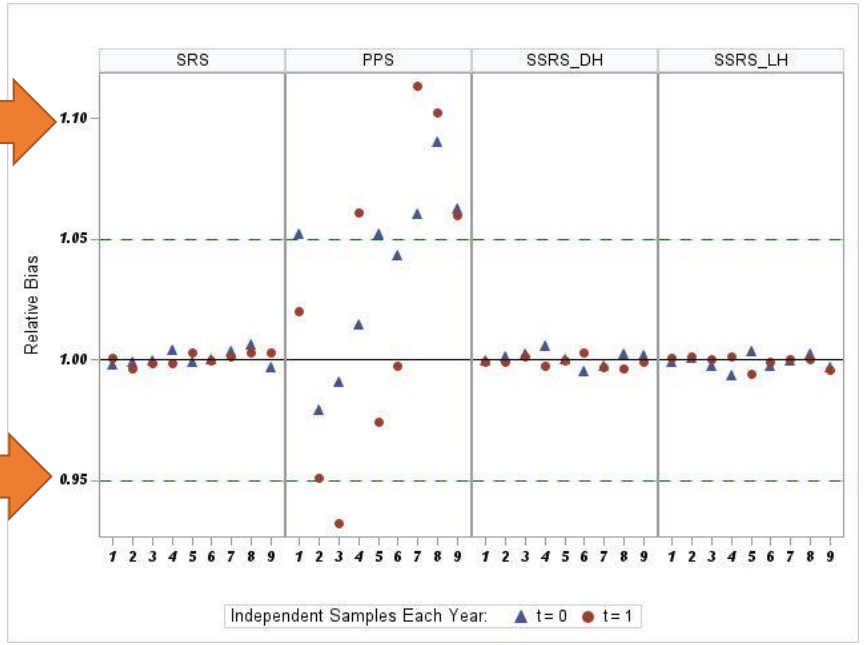
Unit Census Bureau

Relative Bias: "New Survey" (Case 1)

R&D Prevalence (Proportion with R&D)



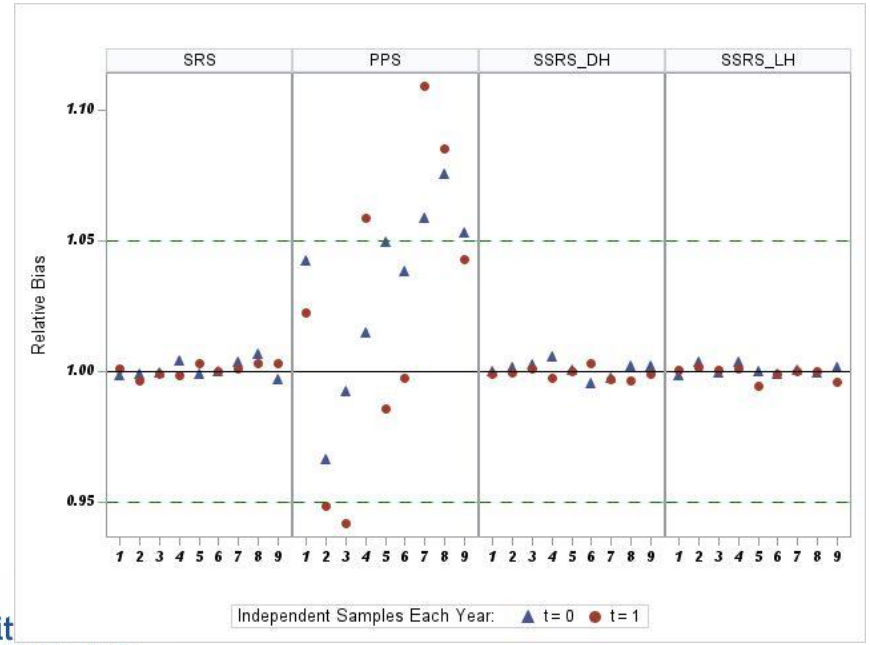
Total R&D Expenditures



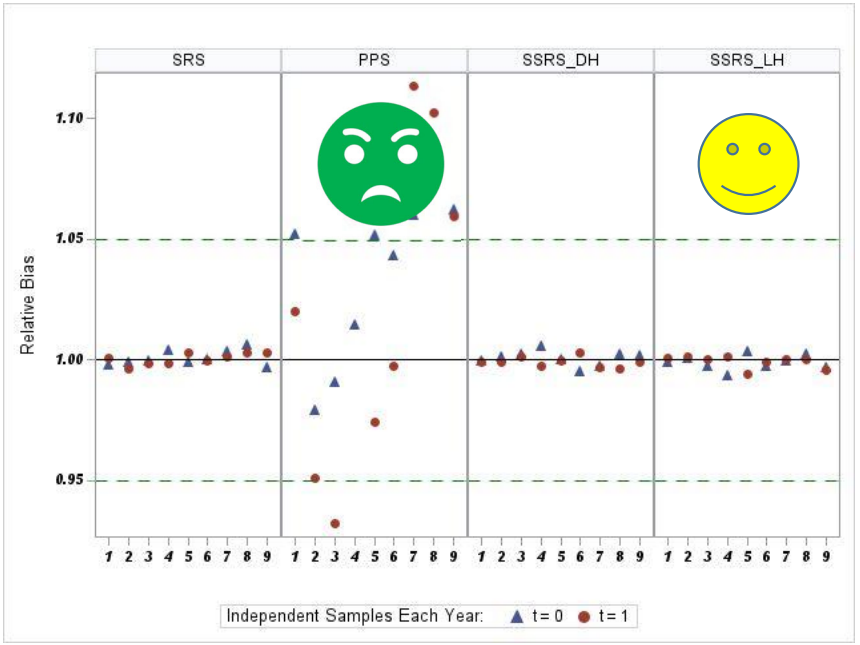
Unit Census Bureau

Relative Bias: "New Survey" (Case 1)

R&D Prevalence (Proportion with R&D)



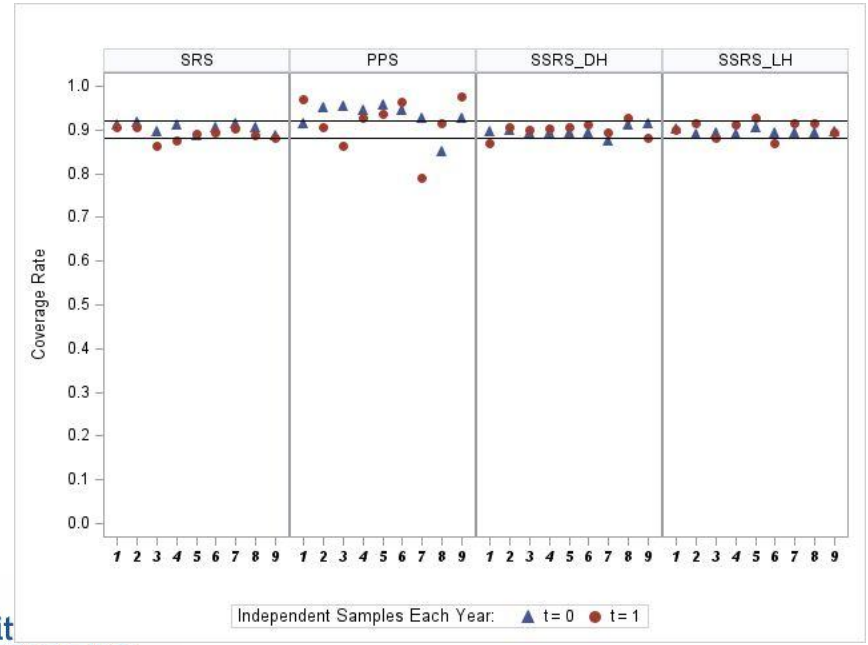
Total R&D Expenditures



Unit Census Bureau

90% CI Coverage Rate: "New Survey"

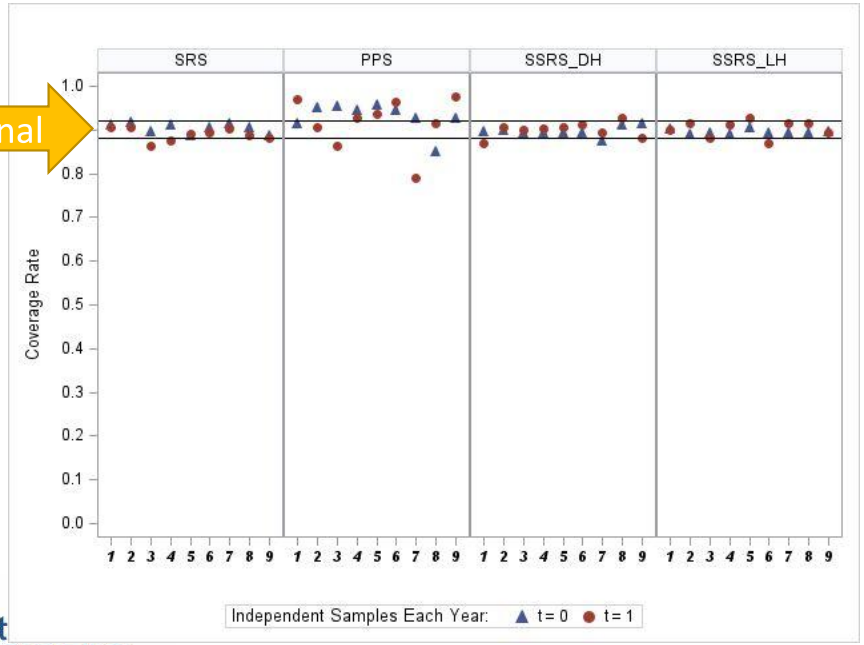
R&D Prevalence (Proportion with R&D)



Unit Census Bureau

90% CI Coverage Rate: "New Survey"

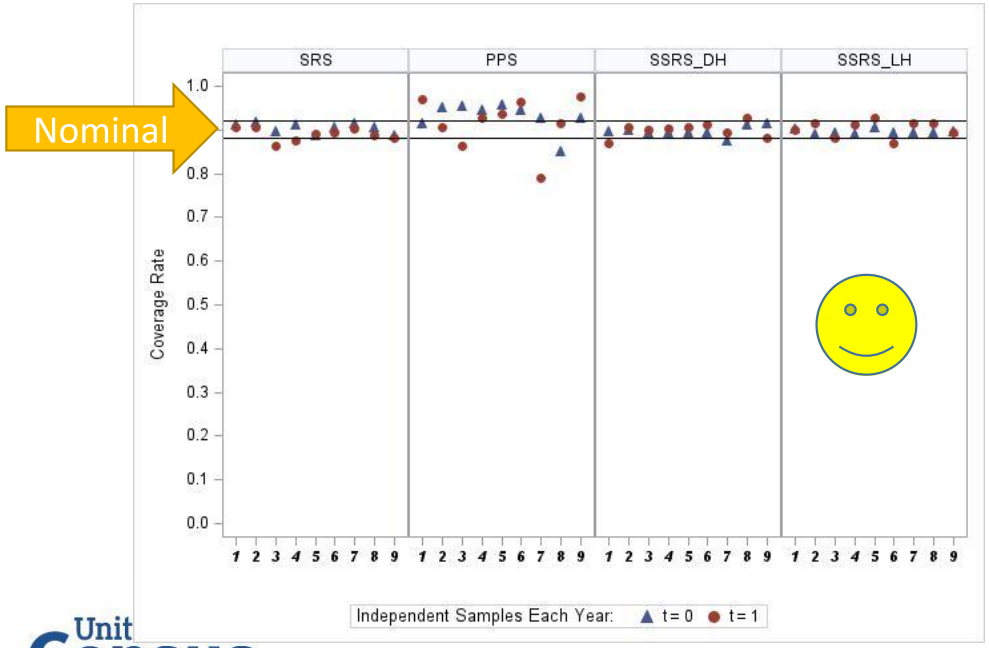
R&D Prevalence (Proportion with R&D)



Unit Census Bureau

90% CI Coverage Rate: "New Survey"

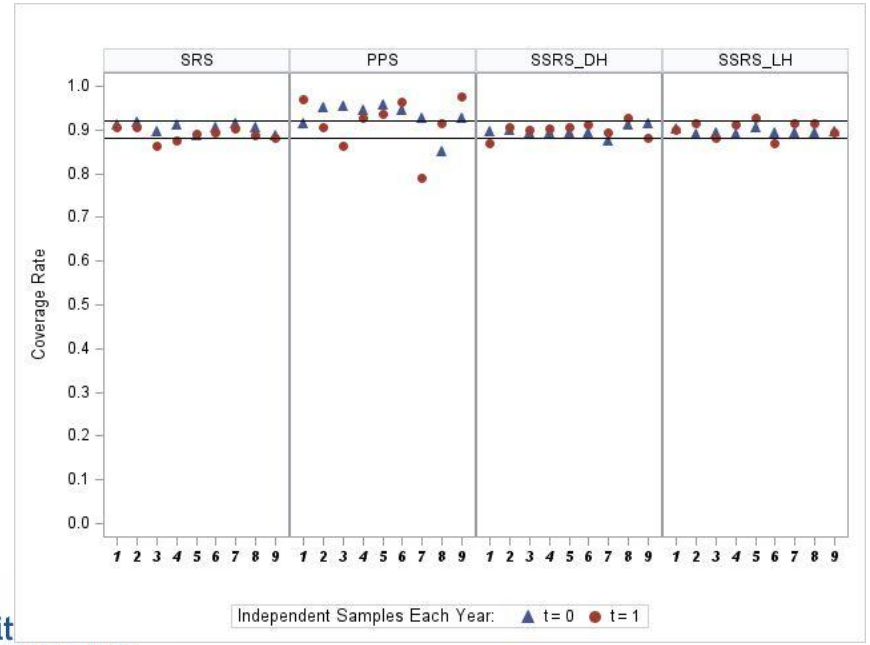
R&D Prevalence (Proportion with R&D)



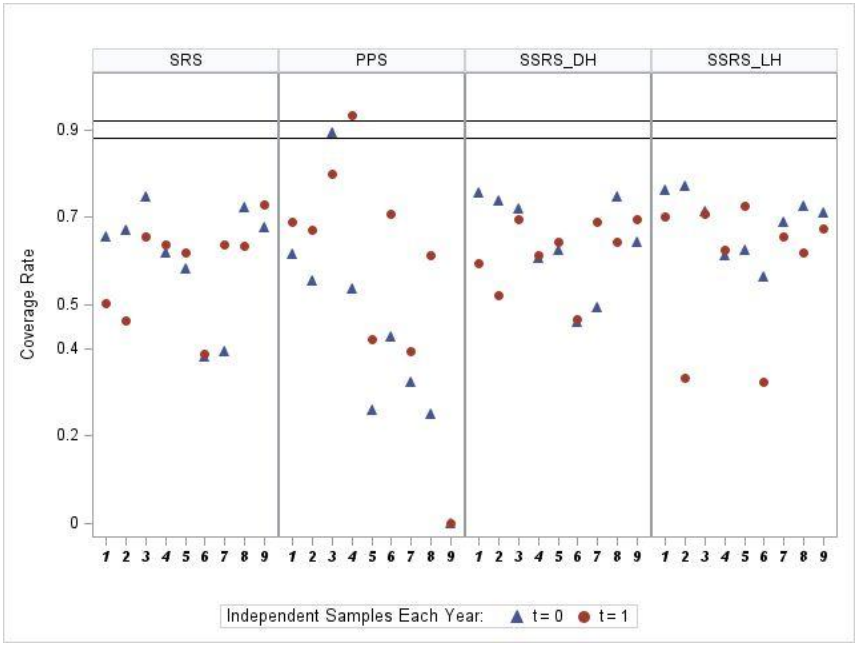
Unit Census Bureau

90% CI Coverage Rate: "New Survey"

R&D Prevalence (Proportion with R&D)



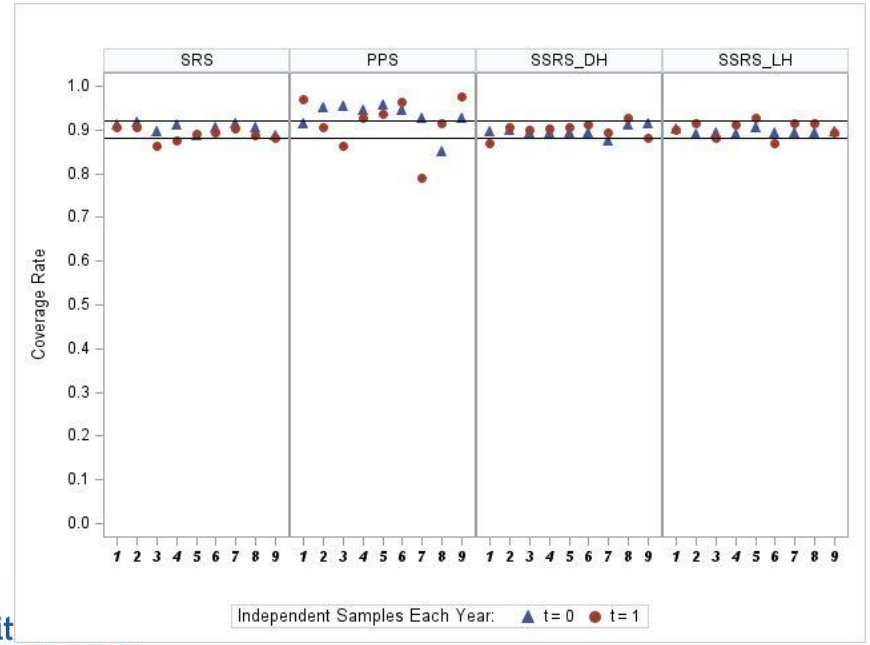
Total R&D Expenditures



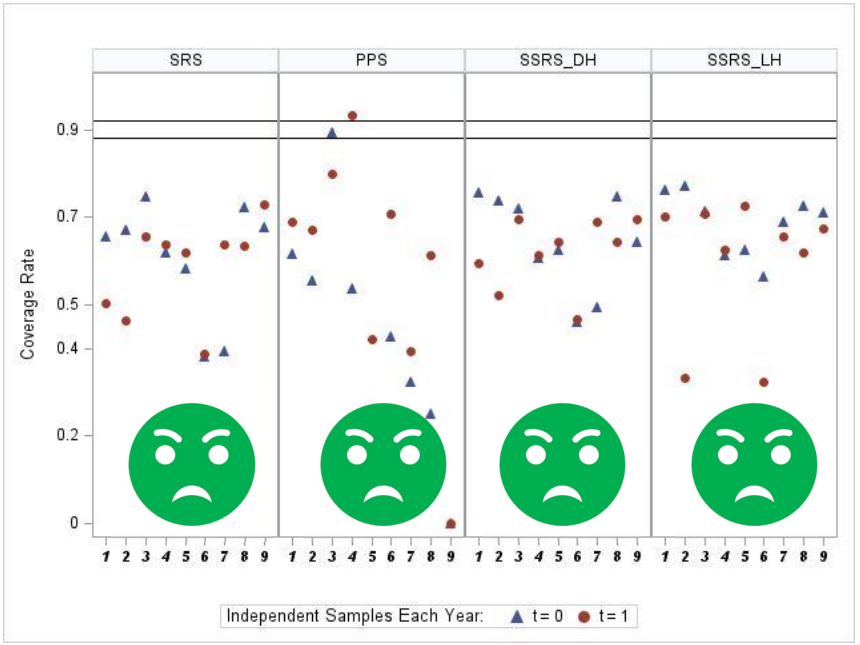
Unit Census Bureau

90% CI Coverage Rate: "New Survey"

R&D Prevalence (Proportion with R&D)



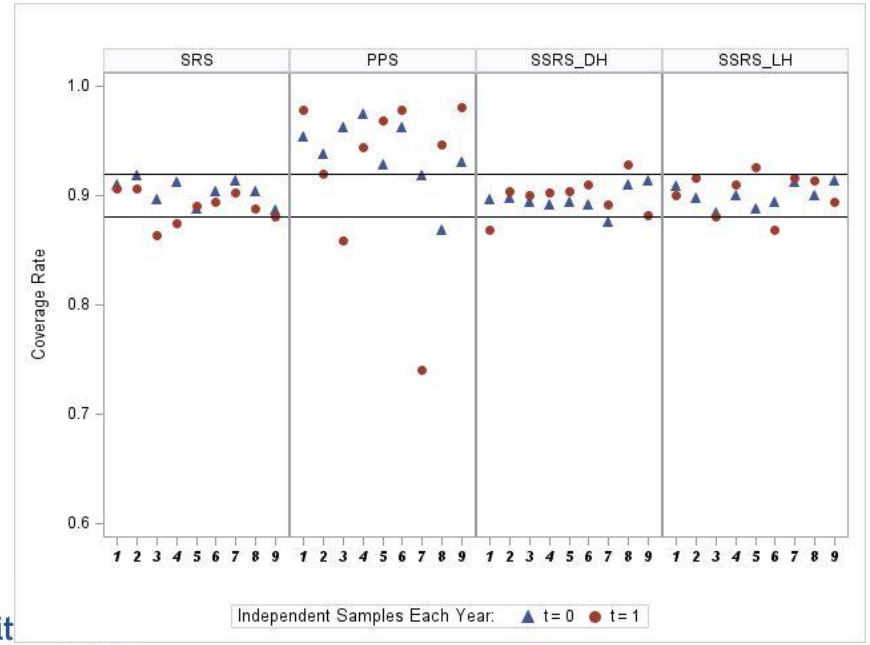
Total R&D Expenditures



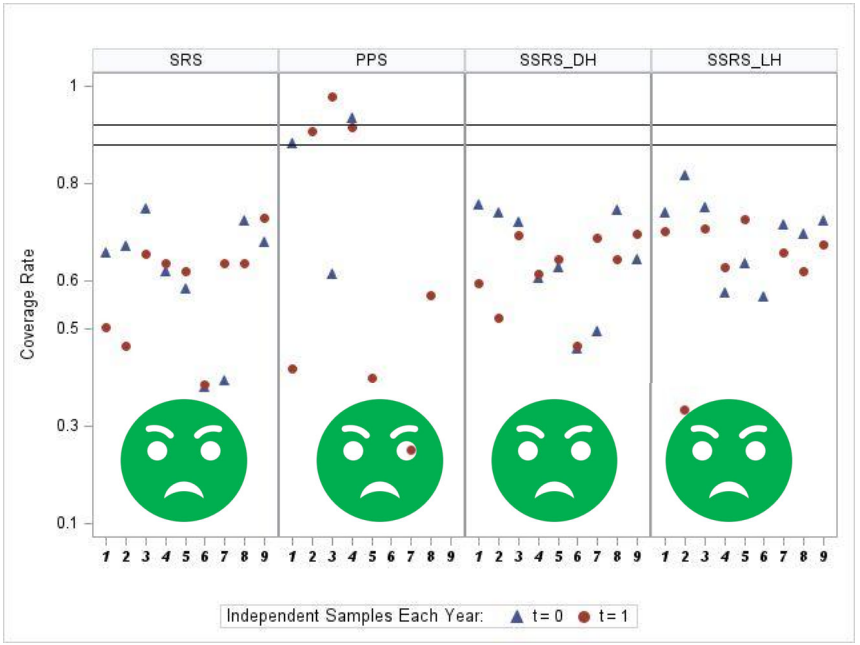
Unit Census Bureau

90% CI Coverage Rate: "New Survey"

R&D Prevalence (Proportion with R&D)



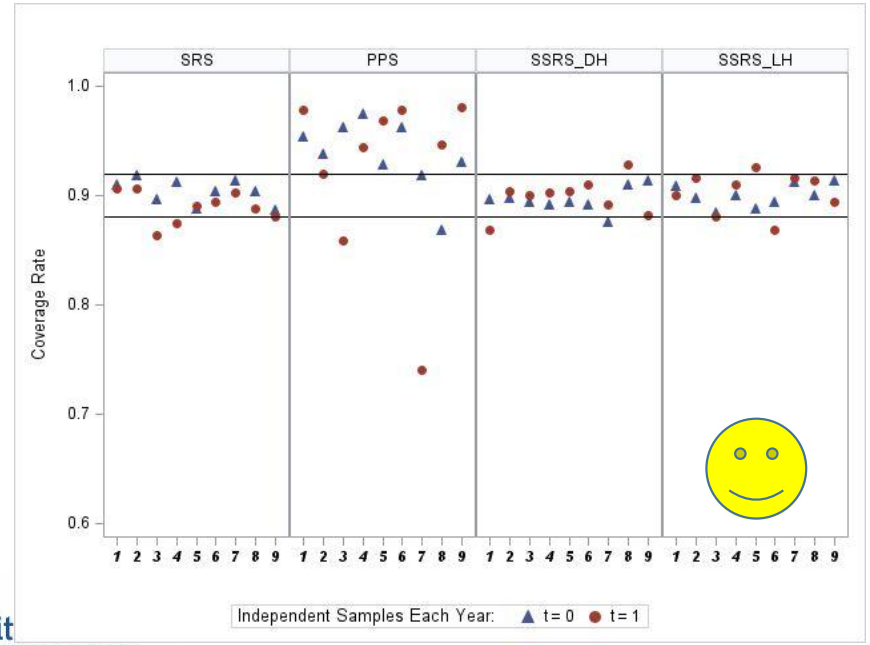
Total R&D Expenditures



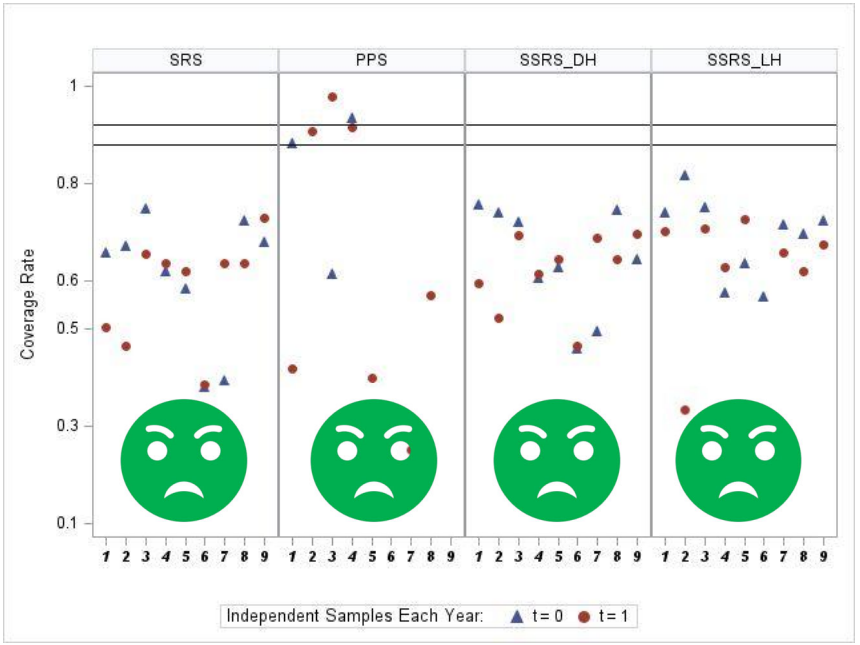
Unit Census Bureau

90% CI Coverage Rate: "New Survey"

R&D Prevalence (Proportion with R&D)

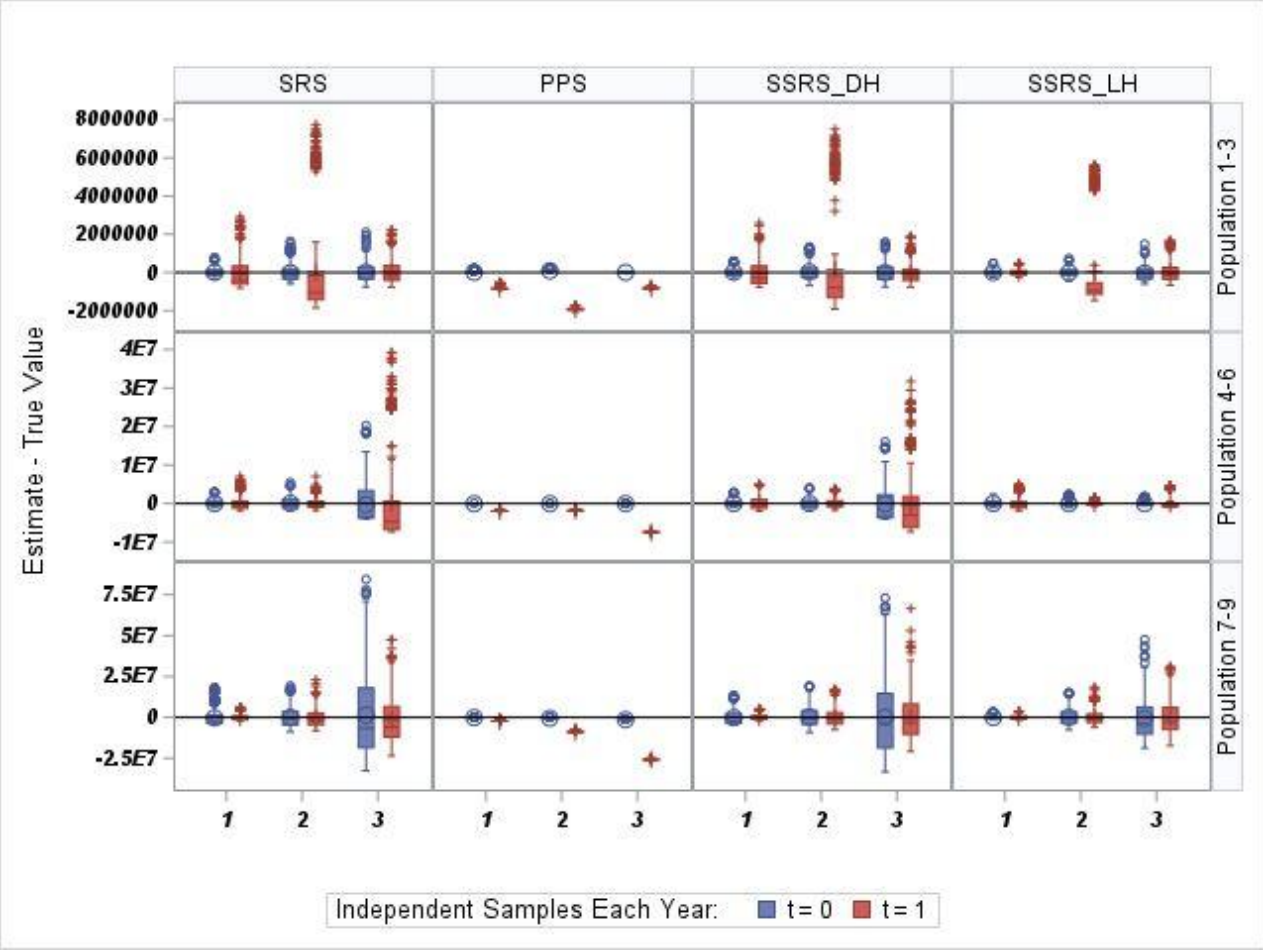


Total R&D Expenditures

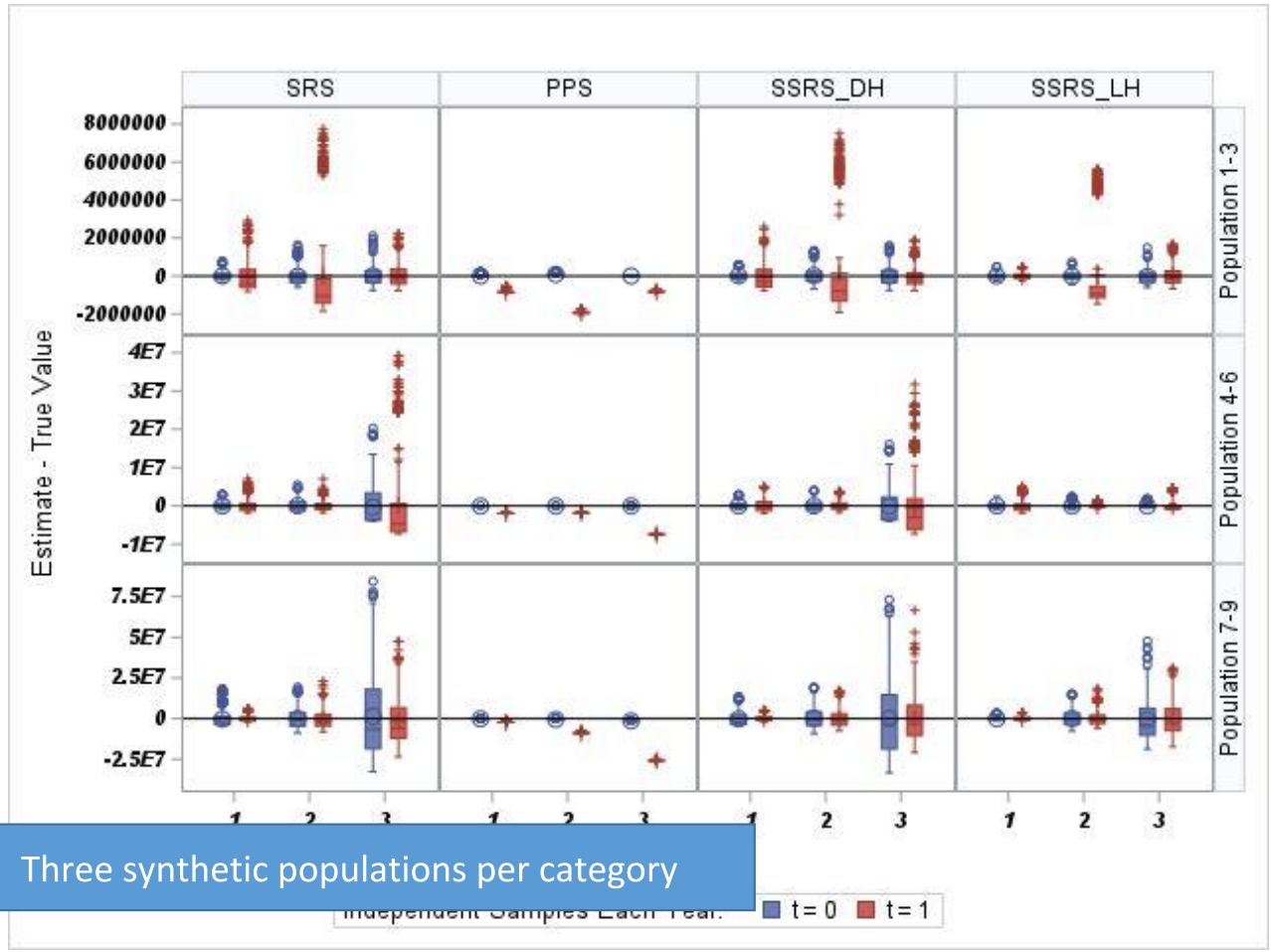


Unit Census Bureau

Let's Dive Deeper…Total R&D Expenditures



Let's Dive Deeper…Total R&D Expenditures



R&D in smaller units

R&D throughout population

R&D in larger units

Three synthetic populations per category

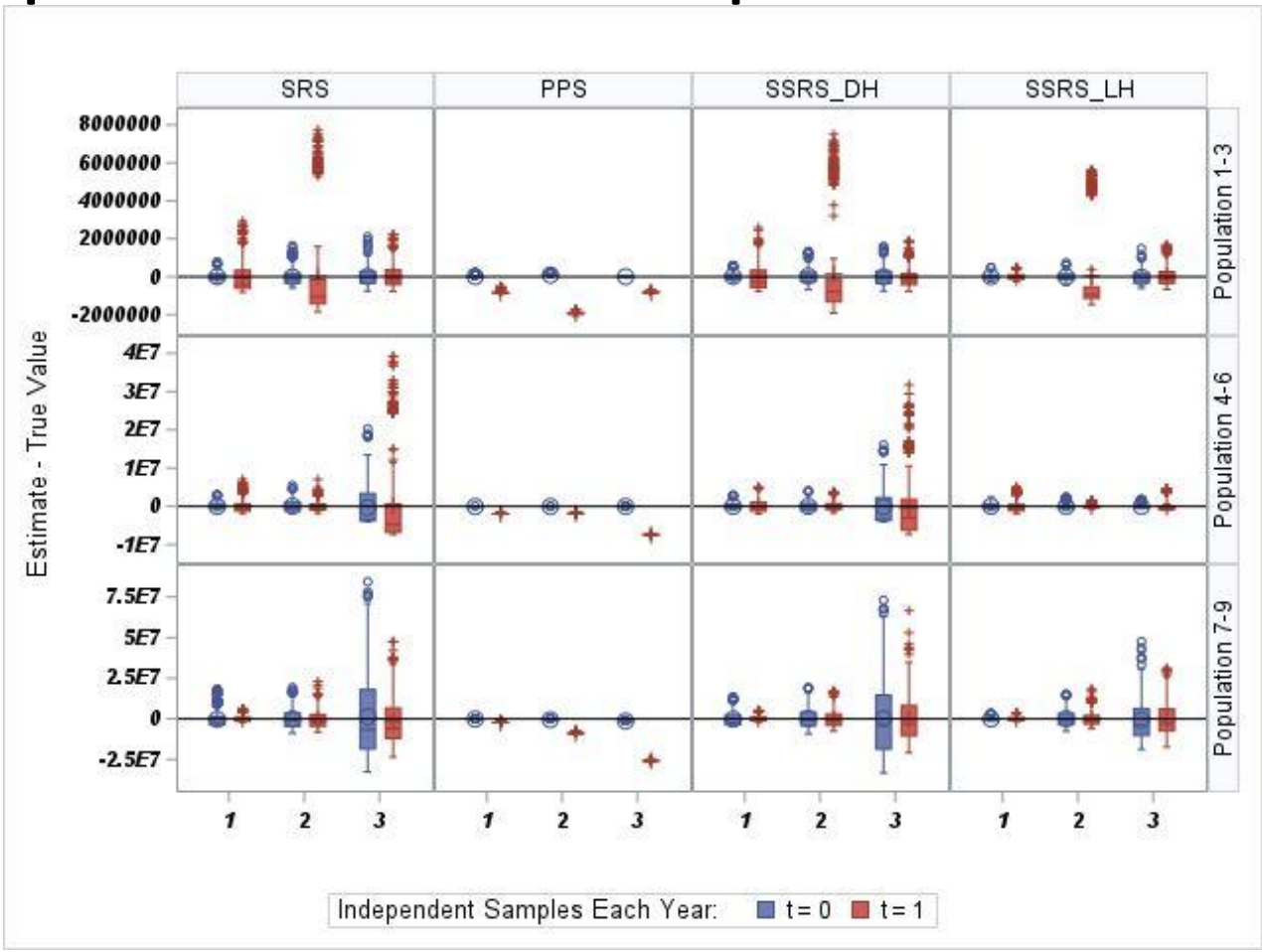
United States® Census Bureau

Let's Dive Deeper…Total R&D Expenditures

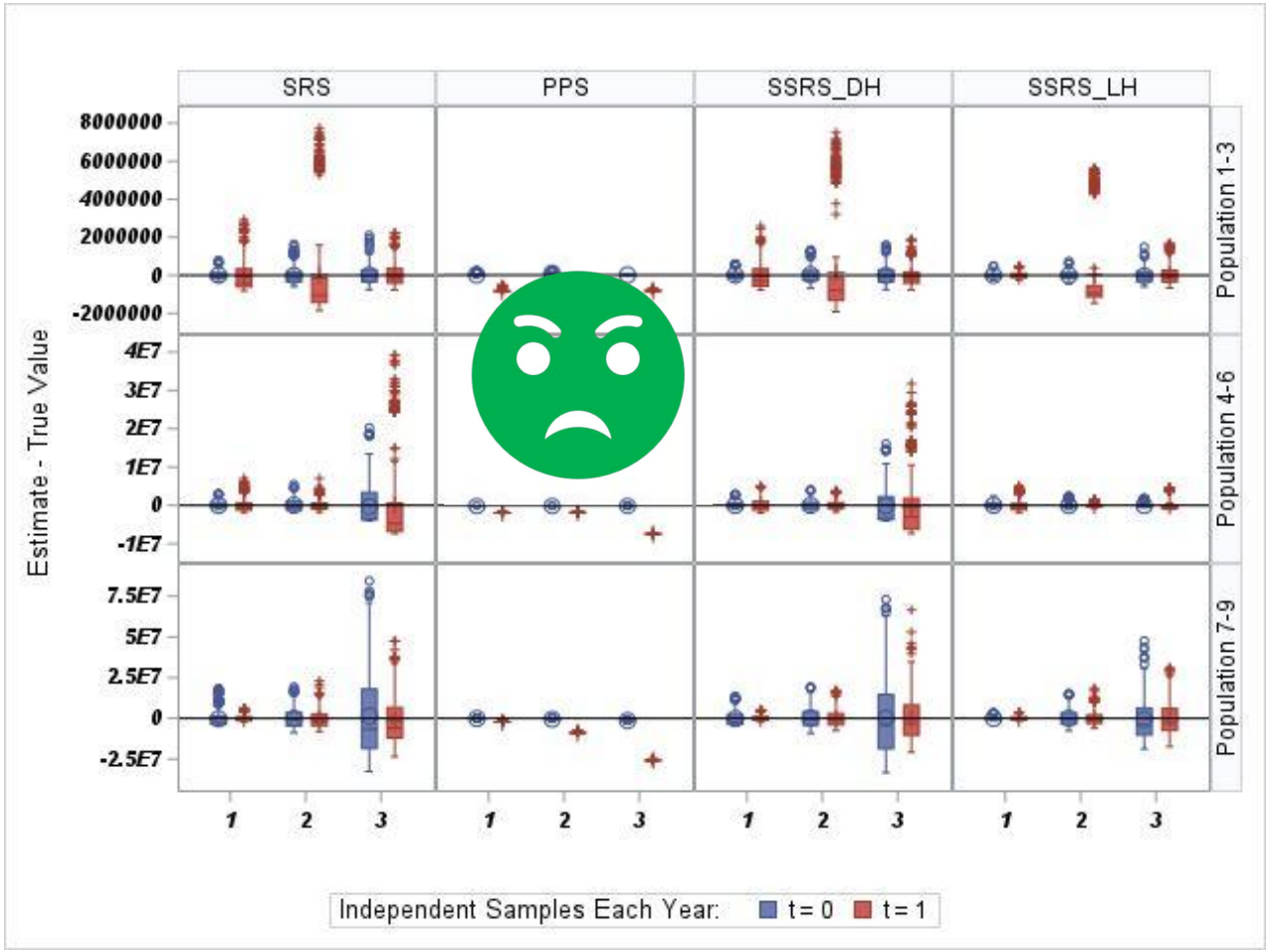
In each sample (within design), subtract true value from estimate.

Ideally,

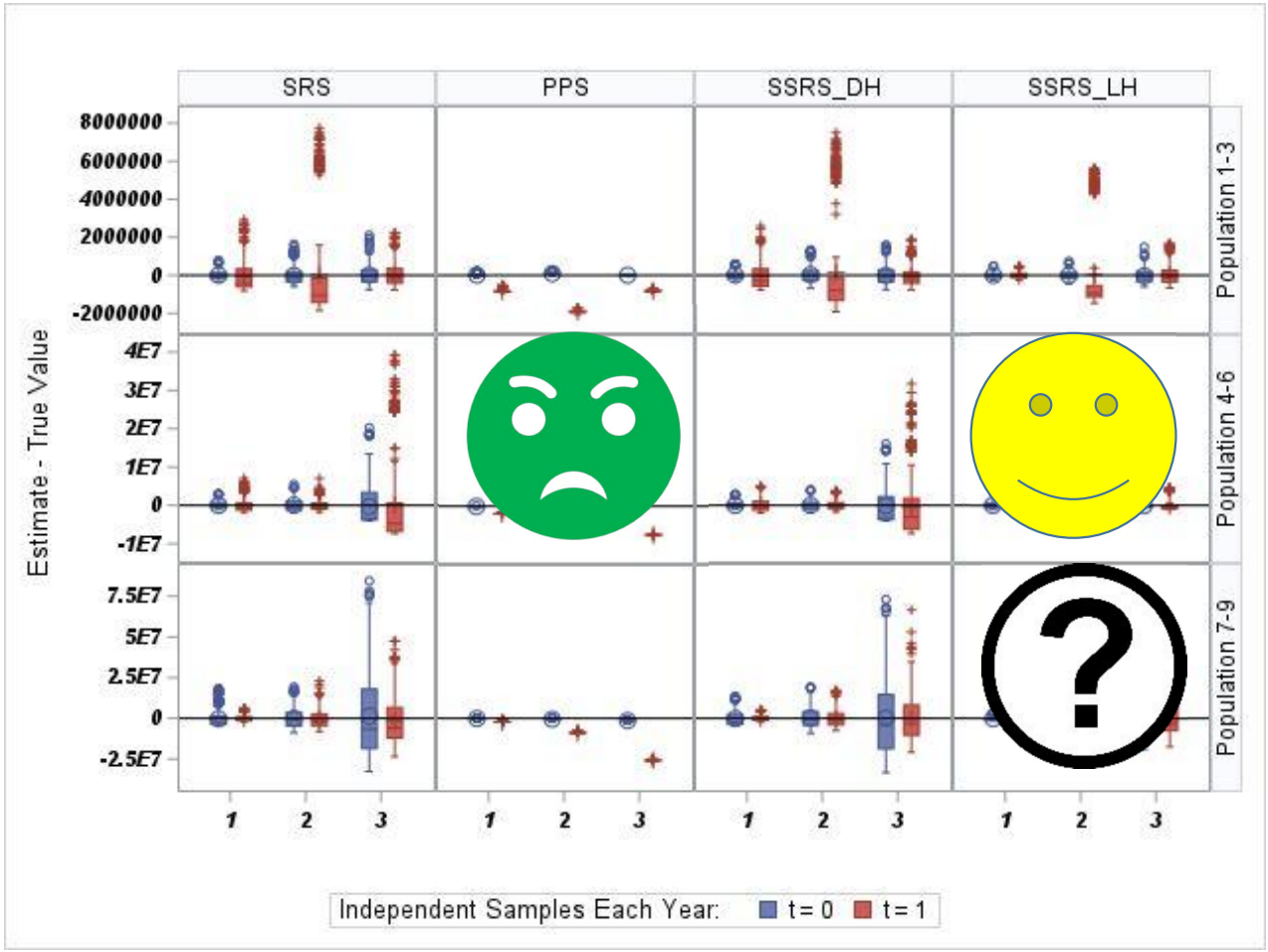
1. Mean = 0 (unbiased)
2. Spread = symmetric



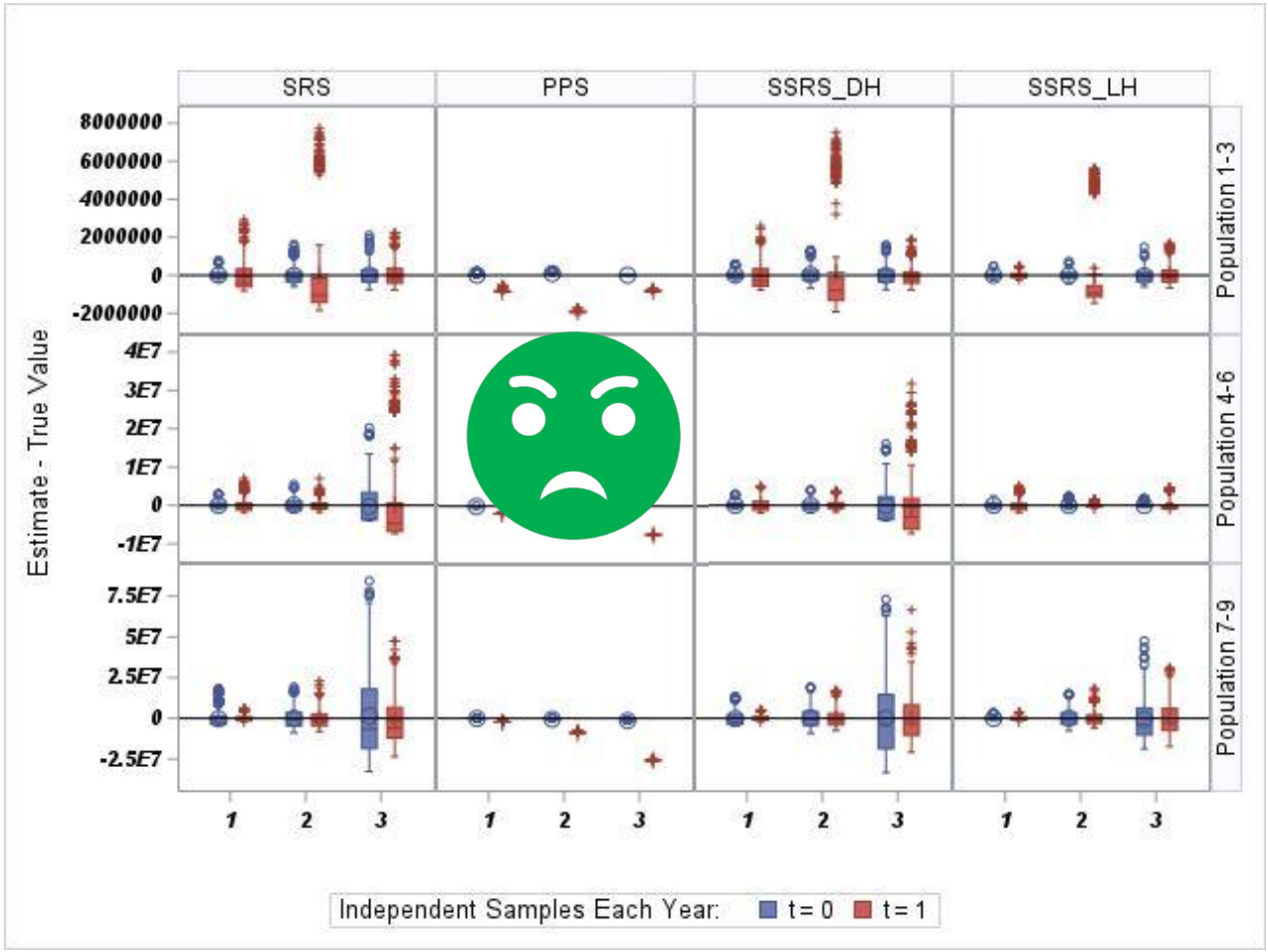
Let's Dive Deeper…Total R&D Expenditures



Let's Dive Deeper…Total R&D Expenditures

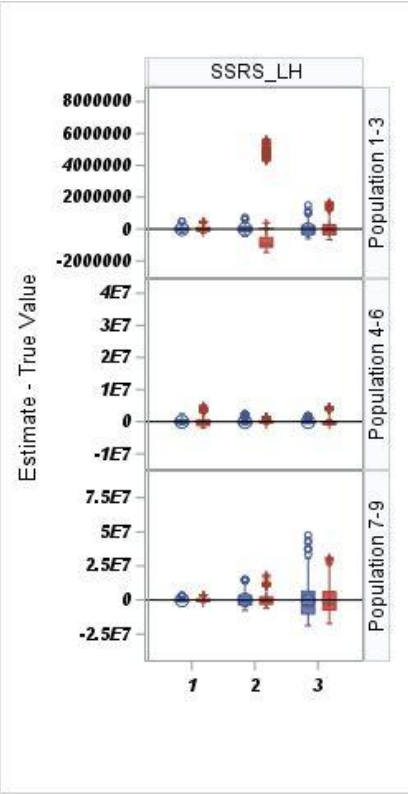


Let's Dive Deeper…Total R&D Expenditures



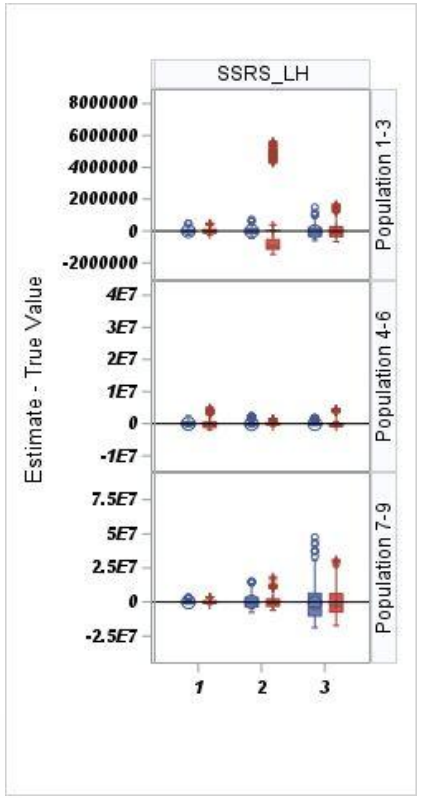
United States® Census Bureau

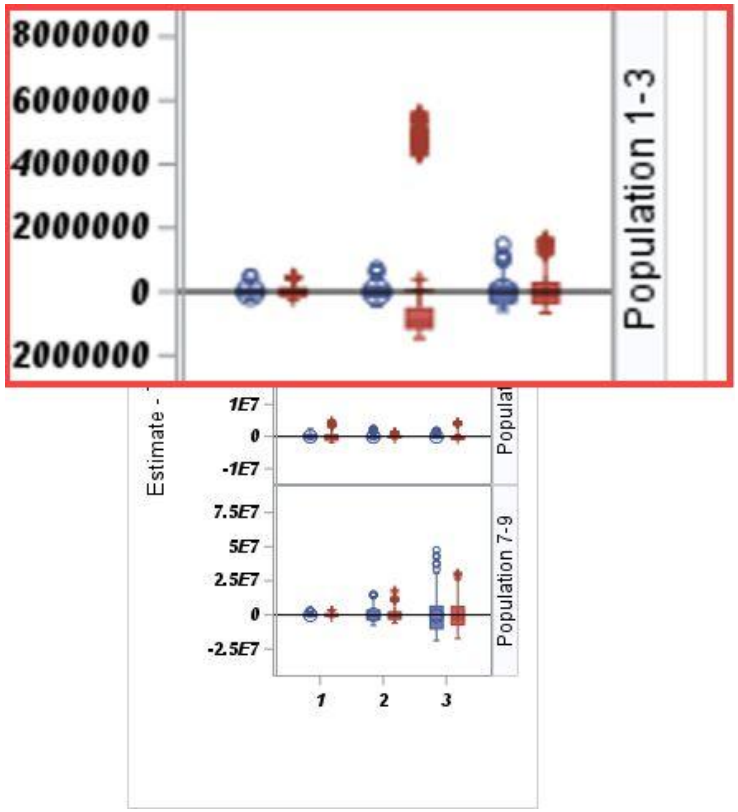
Let’s Keep Going…



United States® Census Bureau

Let's Keep Going…



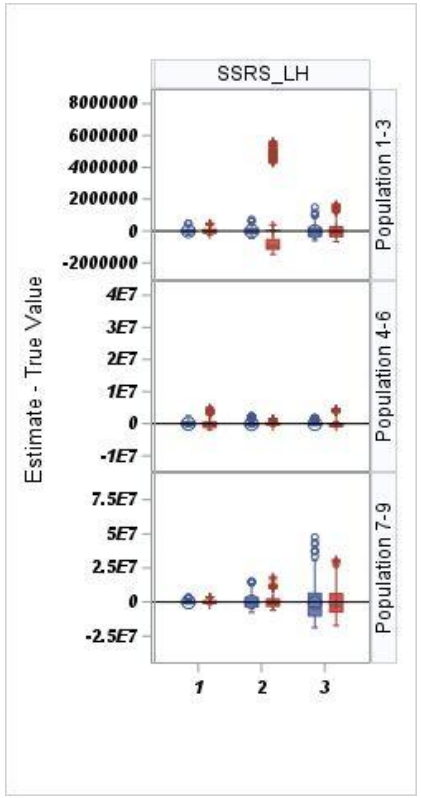


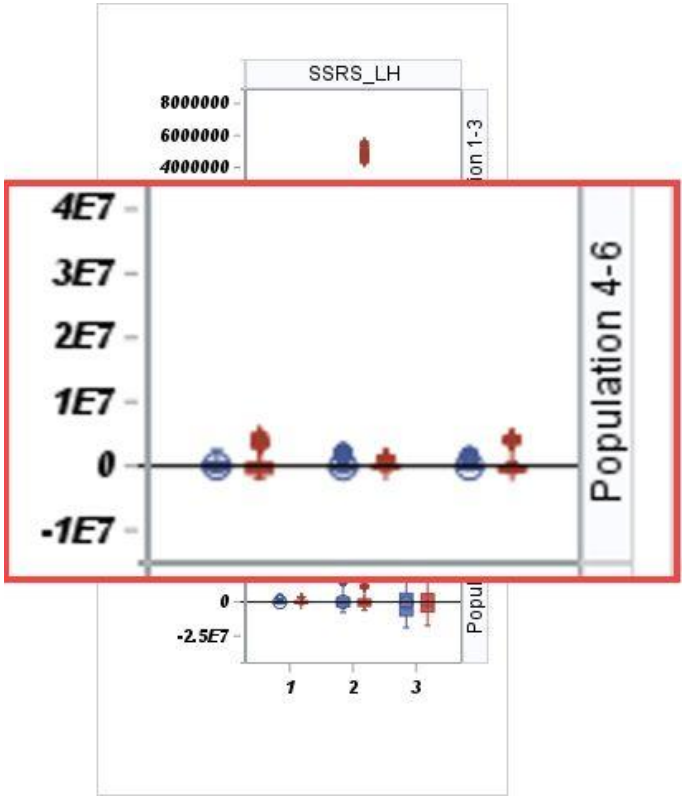
- Right skewed…
- ≈25% of the samples are overestimates
- Severe underestimation in population 2 (*t =1*)
- Precise estimates (small c.v.s)

Undercoverage!

United States® Census Bureau

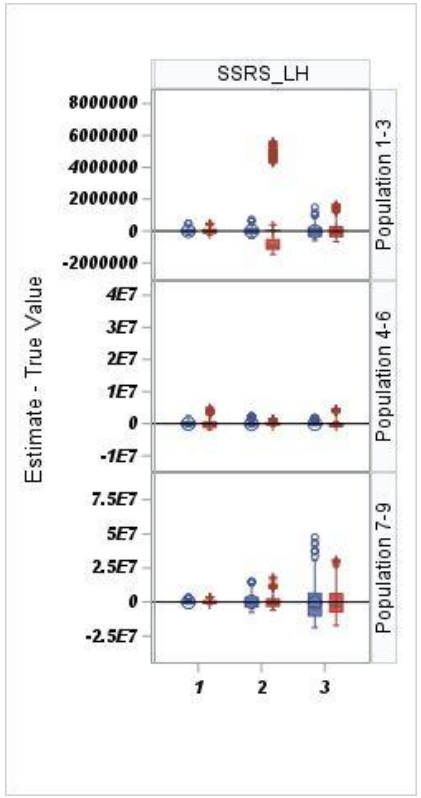
Let’s Keep Going…

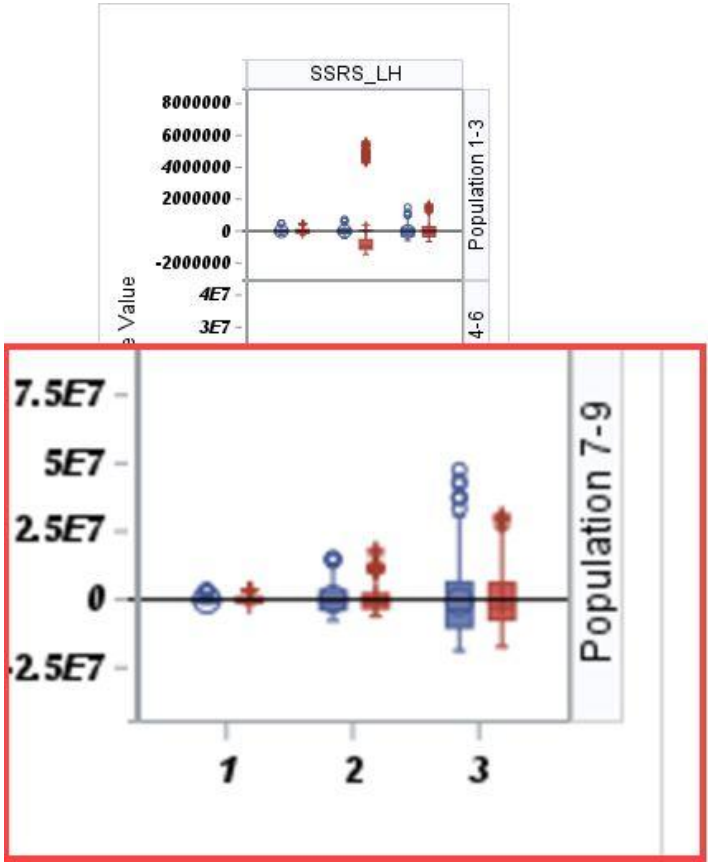




Same Story…

Let's Keep Going…





You knew this was coming!

United States® Census Bureau

Where Are We?

- Evidence for using size-based stratification in sample design
  - Improvements in relative bias of total R&D expenditures over other methods

- Evidence against unequal probability sampling
  - Annual payroll as measure of size

- BIG CAUTION
  - Basing sample design recommendation based entirely on c.v. results seems unwise (can get very precise but quite biased estimates)

United States® Census Bureau

Moreover,

- There is severe undercoverage for total R&D expenditures
  - All sample designs
  - All unit size/propensity combinations

- Can we modify any sample designs to improve the 90% coverage rates for total R&D expenditures?

United States® Census Bureau

Can We Incorporate Previous Sample Results Into Current Sample Design?

	Candidate Sample Design	Independent Sample	Use Previous Sample
SRS	Simple random sample without replacement	No stratification	
PPS	Pareto sample without replacement	No stratification	
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	6 strata per industry	
SSRS_LH	Stratified SRS, Lavallée Hidiroglou (LH) method and Dalenius Hodges (DH)	1 certainty stratum 5 noncertainty strata	

United States® Census Bureau

Can We Incorporate Previous Sample Results Into Current Sample Design?

	Candidate Sample Design	Independent Sample	Use Previous Sample
SRS	Simple random sample	No stratification	
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	6 strata per industry	
SSRS_LH	Stratified SRS, Lavallée Hidiroglou (LH) method and Dalenius Hodges (DH)	1 certainty stratum 5 noncertainty strata	

1. Include previously sampled units that have R&D included with certainty in the new sample.

United States® Census Bureau

Can We Incorporate Previous Sample Results Into Current Sample Design?

	Candidate Sample Design	Independent Sample	Use Previous Sample
1. Include previously sampled units that have R&D included with certainty in the new sample. 2. Reduce assigned (noncertainty) allocation.			
	replacement		
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	6 strata per industry	
SSRS_LH	Stratified SRS, Lavallée Hidiroglou (LH) method and Dalenius Hodges (DH)	1 certainty stratum 5 noncertainty strata	

United States® Census Bureau

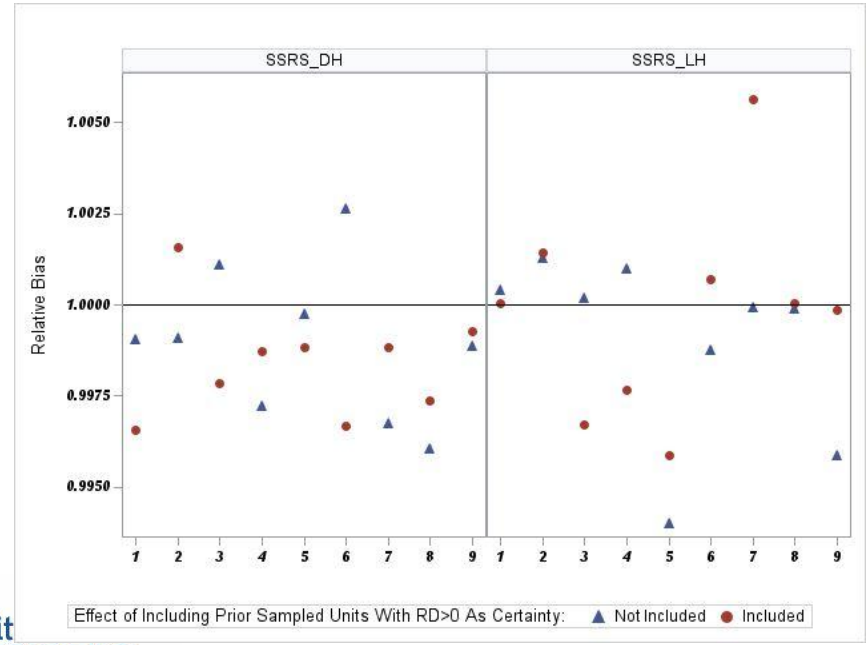
Can We Incorporate Previous Sample Results Into Current Sample Design?

	Candidate Sample Design	Independent Sample	Use Previous Sample
[redacted]			
SSRS_DH	Stratified SRS, Dalenius Hodges (DH) strata	6 strata per industry	Certainty stratum (previous R&D units)
SSRS_LH	Stratified SRS, Lavallée Hidiroglou (LH) method and Dalenius Hodges (DH)	1 certainty stratum 5 noncertainty strata	Add previous R&D units to certainty stratum

United States® Census Bureau

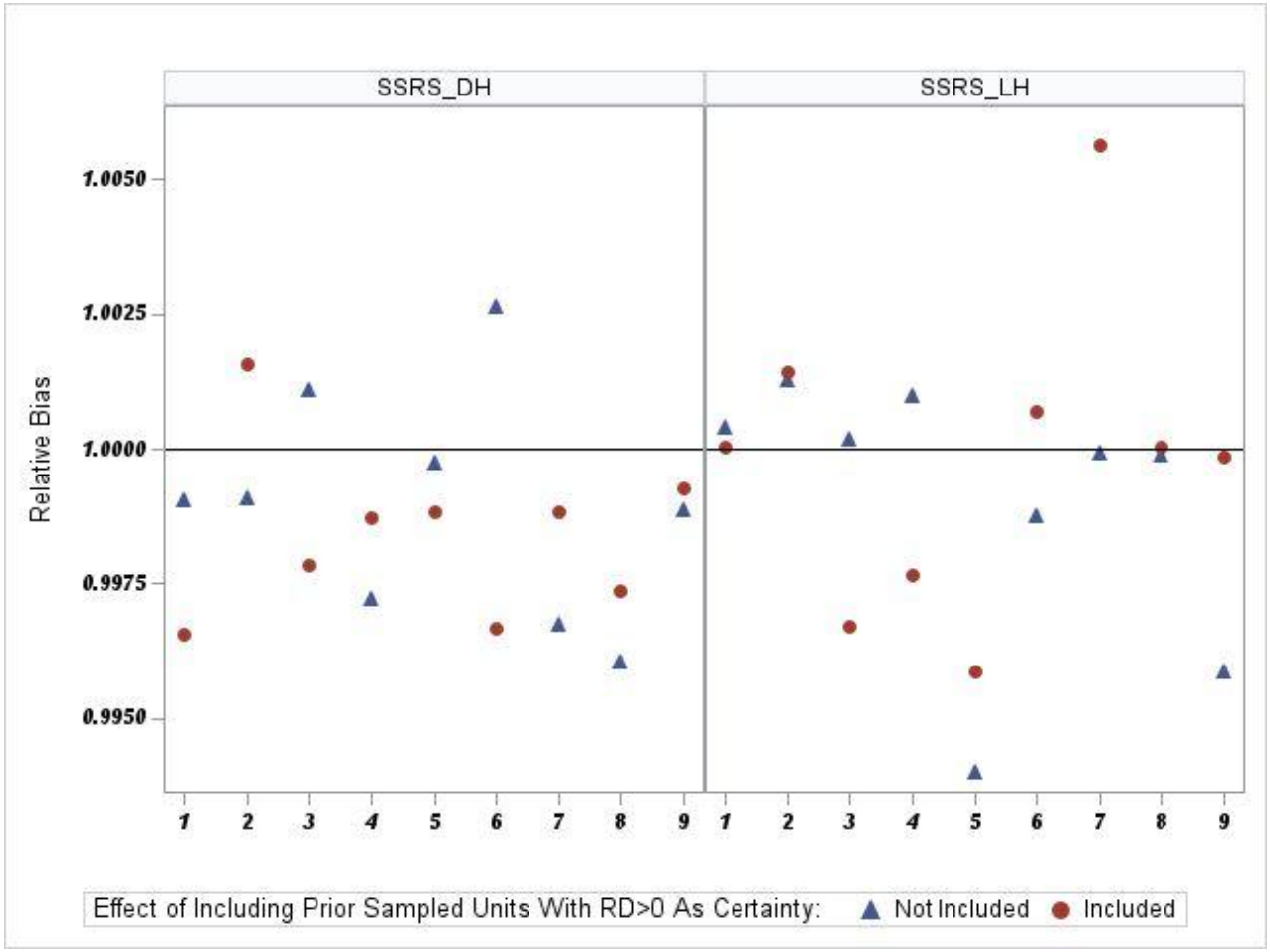
Relative Bias: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D)

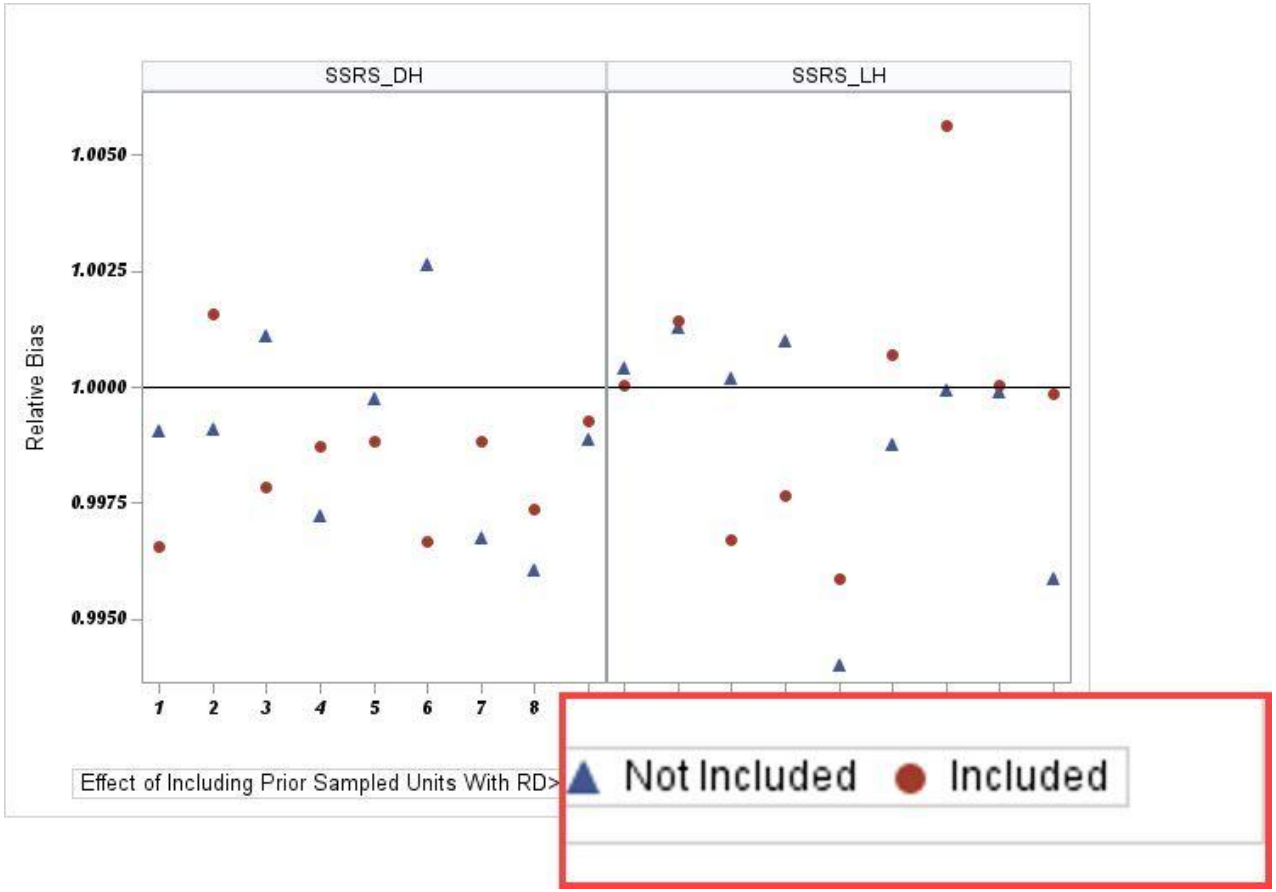


Unit Census Bureau

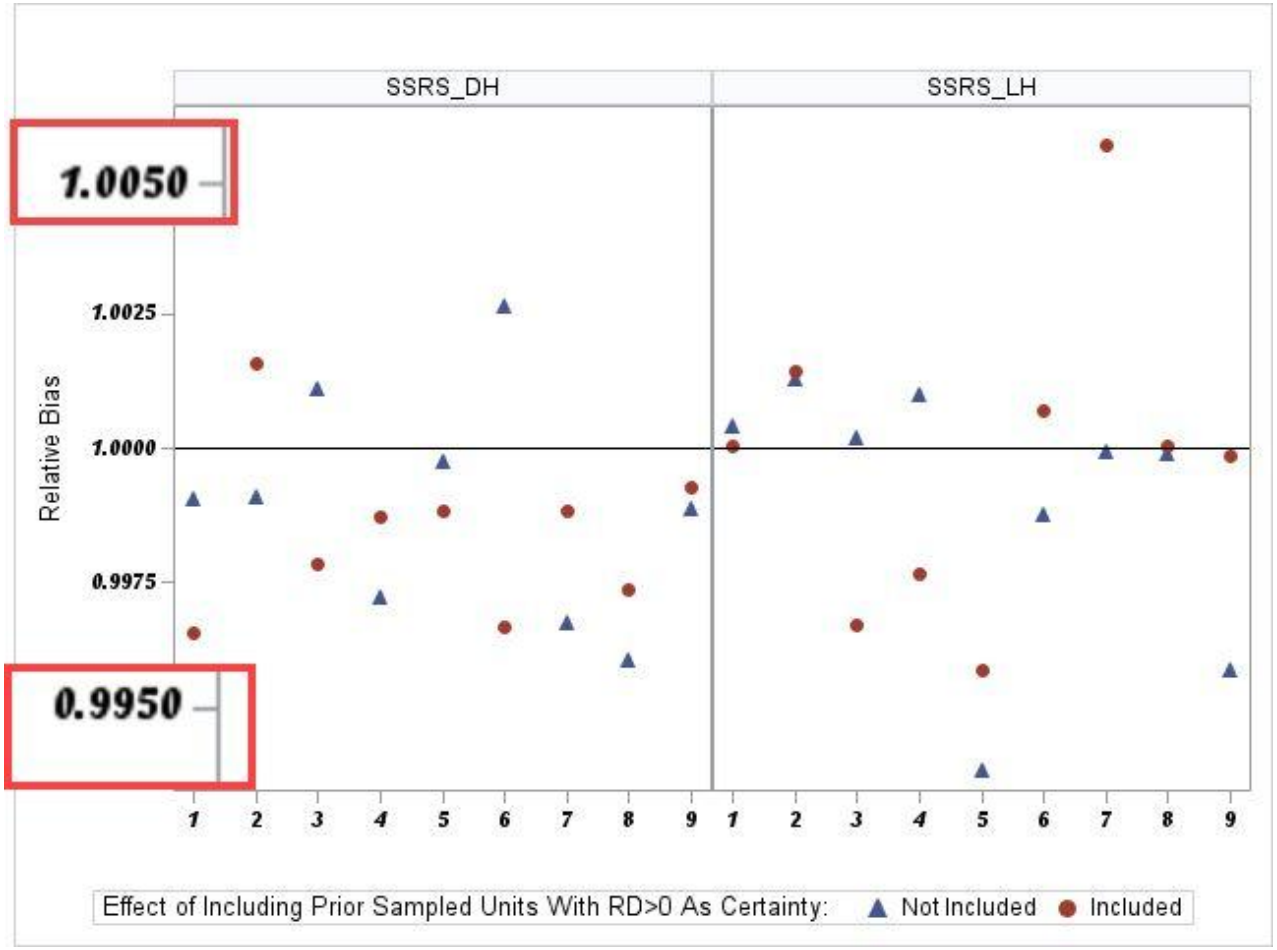
Relative Bias: Prevalence (Case 2)



Relative Bias: Prevalence (Case 2)

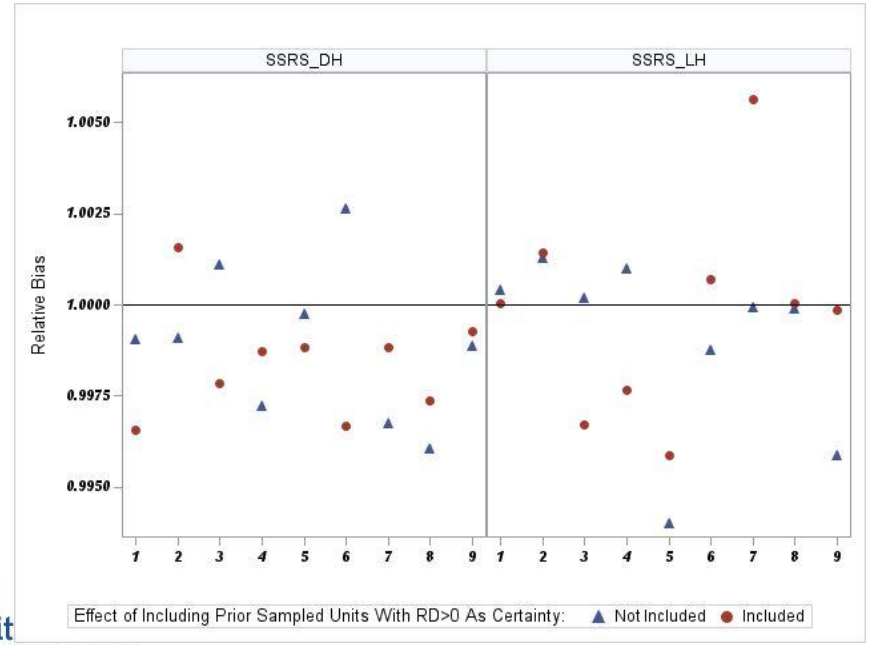


Relative Bias: Prevalence (Case 2)

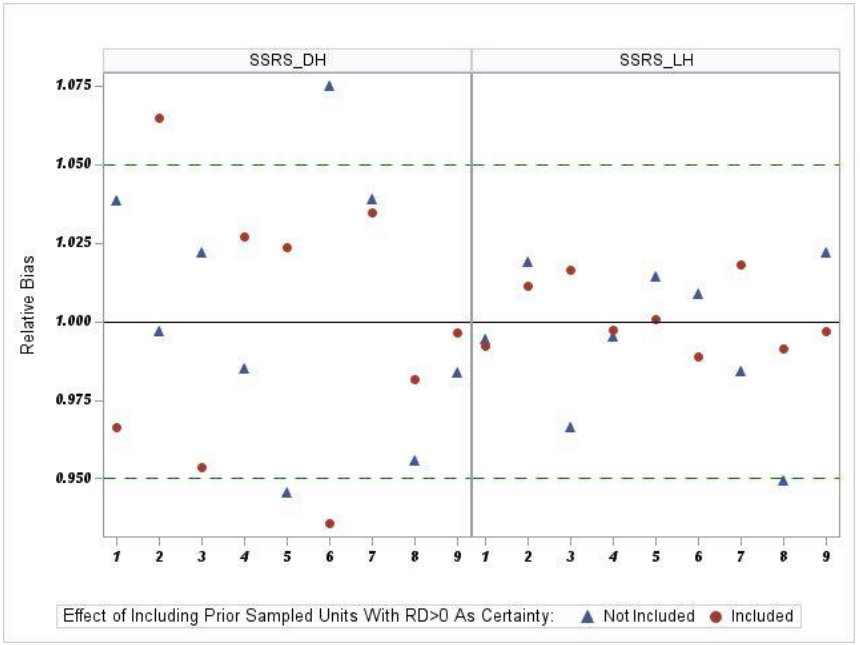


Relative Bias: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D)

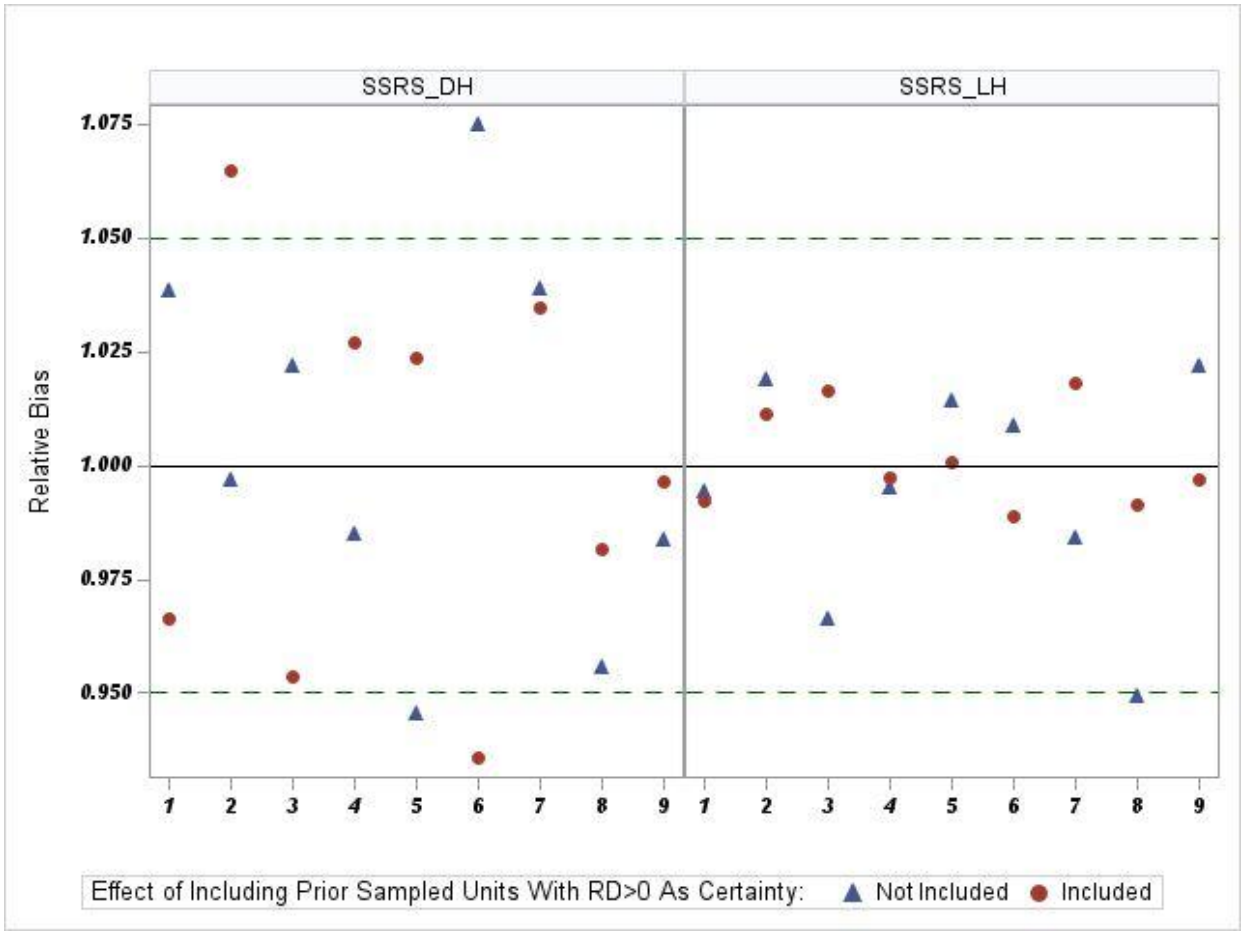


Total R&D Expenditures



Unit Census Bureau

Relative Bias: Total Expenditures (Case 2)

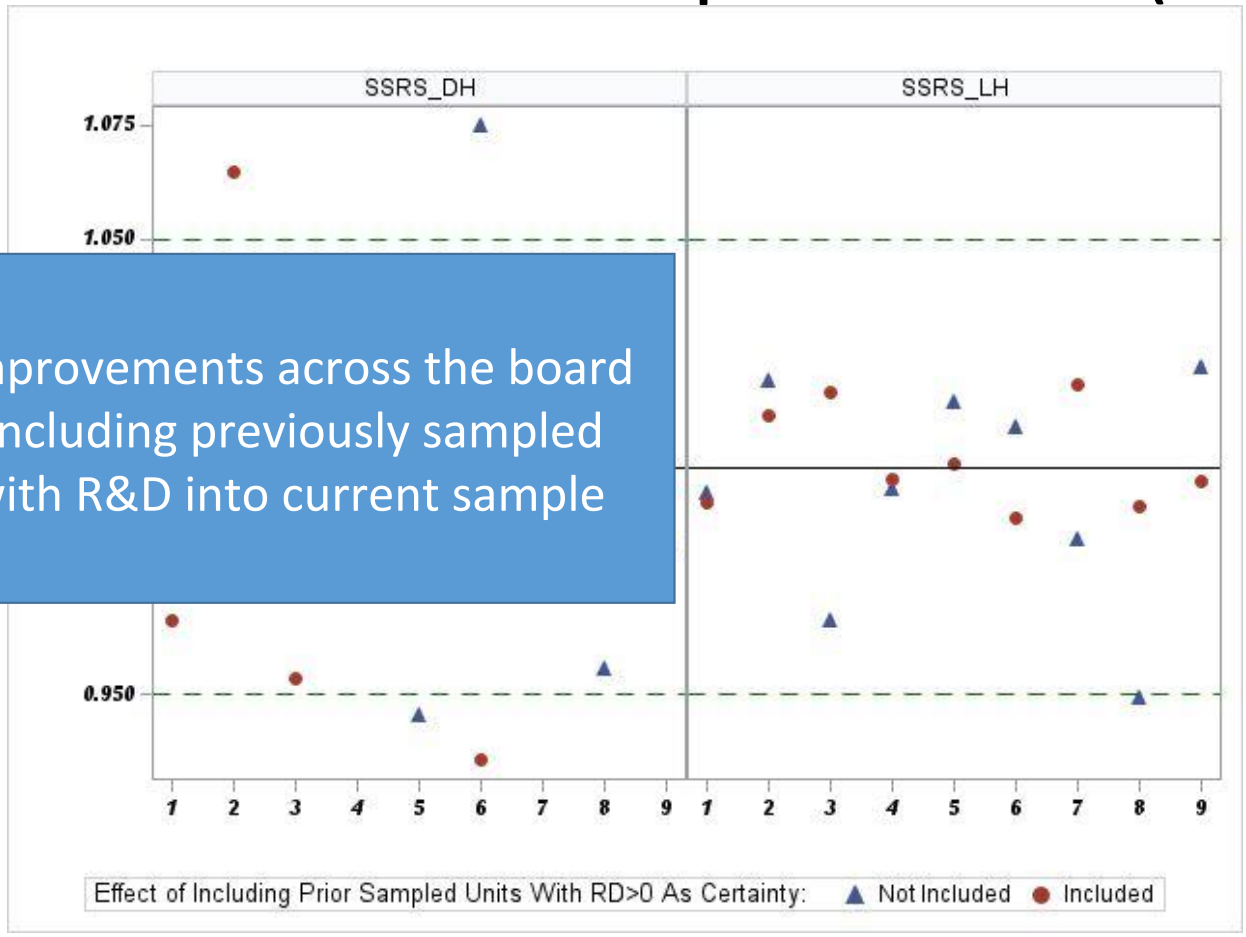


Relative Bias: Total Expenditures (Case 2)

Clear improvements only when larger units tend to expend on R&D

United States® Census Bureau

Relative Bias: Total Expenditures (Case 2)

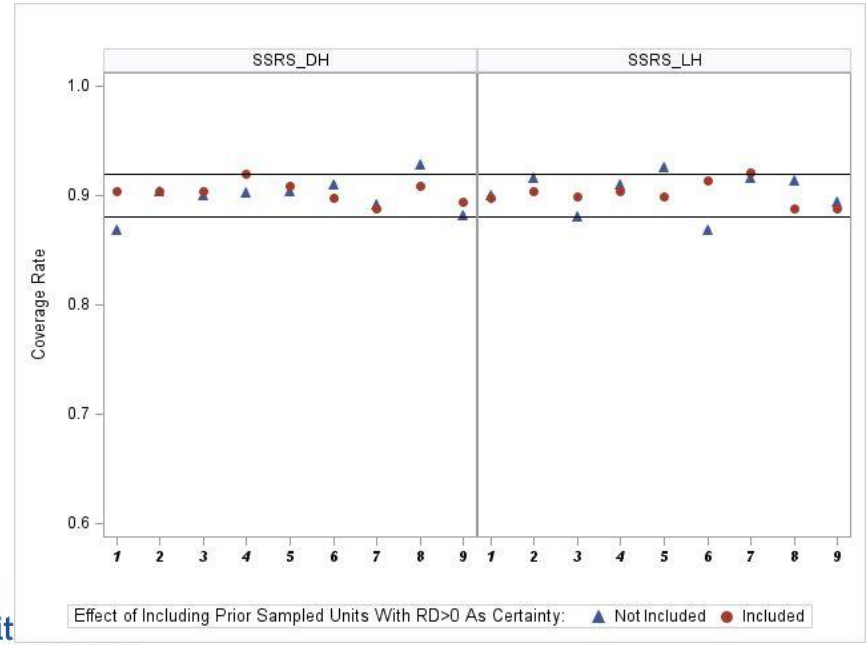


Slight improvements across the board when including previously sampled units with R&D into current sample

United States® Census Bureau

Coverage: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D)

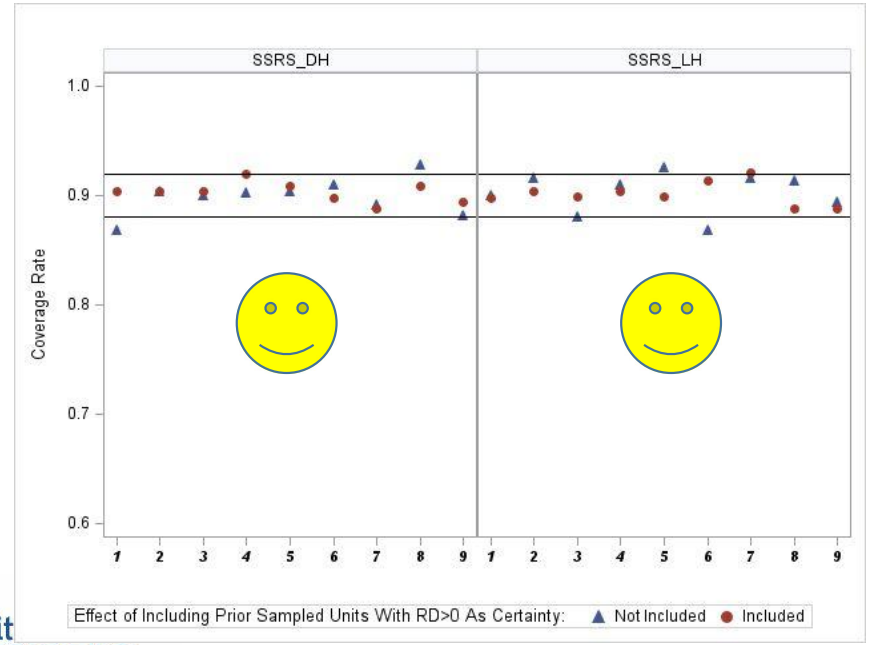


Total R&D Expenditures

Unit Census Bureau

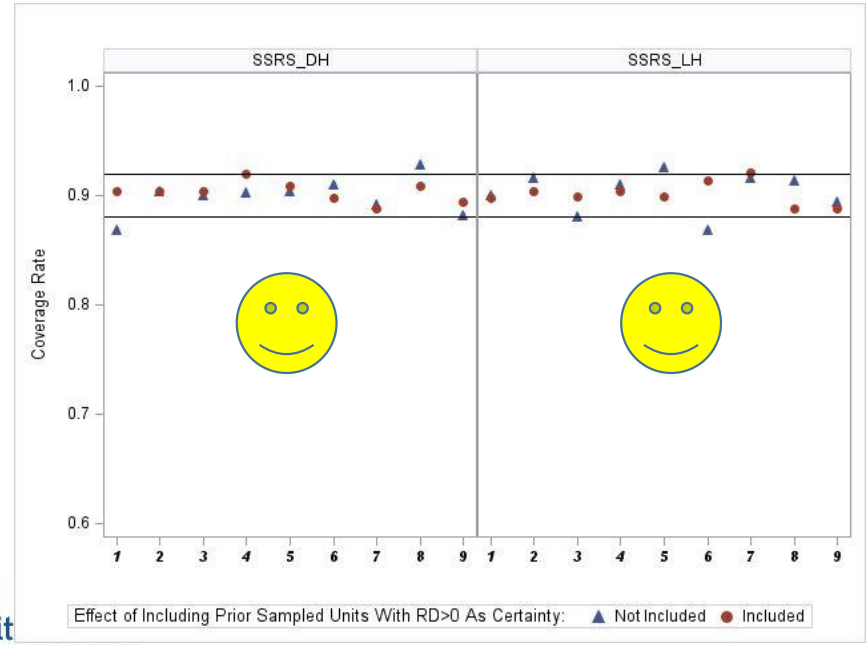
Coverage: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D) **Total R&D Expenditures**



Coverage: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D) **Total R&D Expenditures**

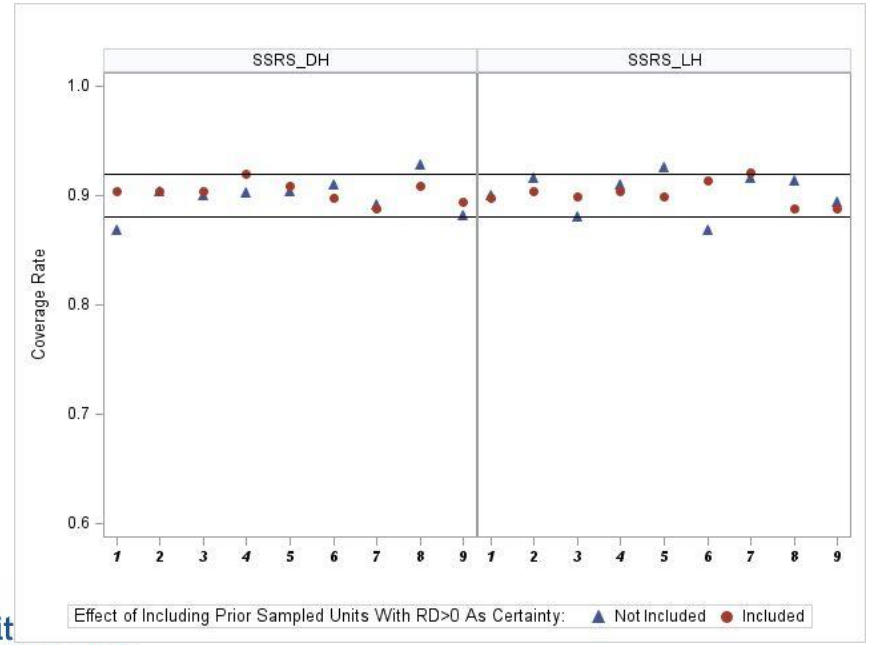


Definite improvements across the board when including previously sample units with R&D into current sample

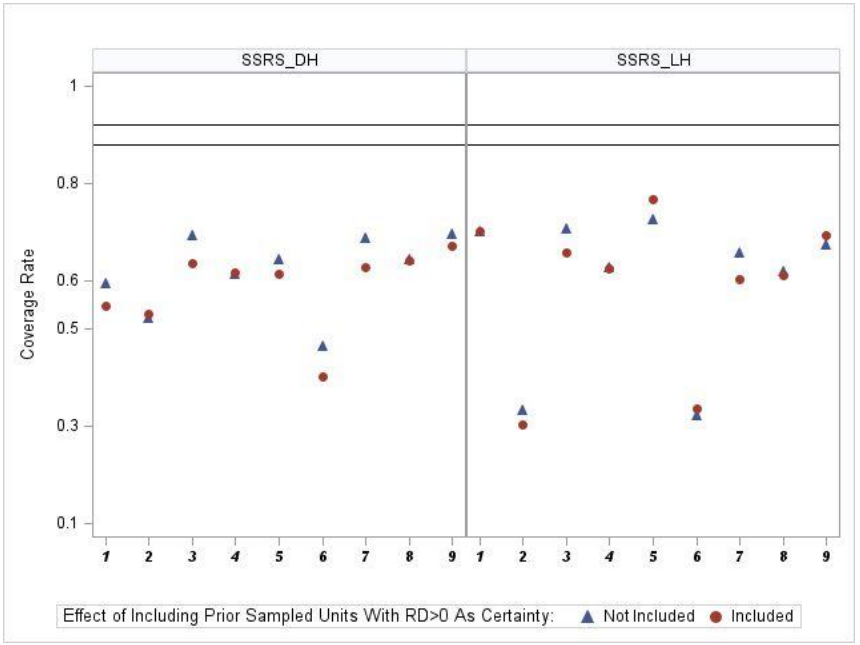
Unit Census Bureau

Coverage: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D)



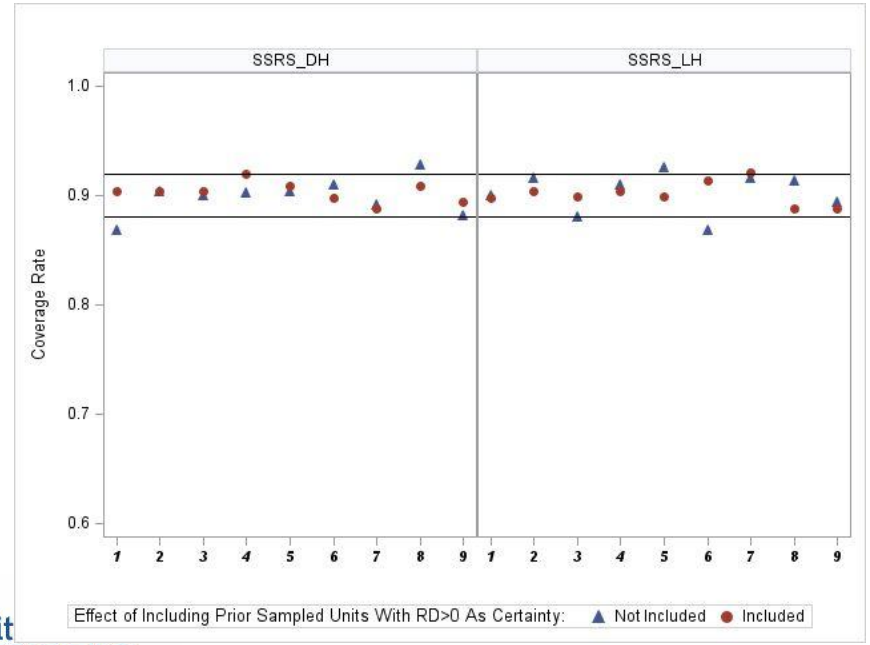
Total R&D Expenditures



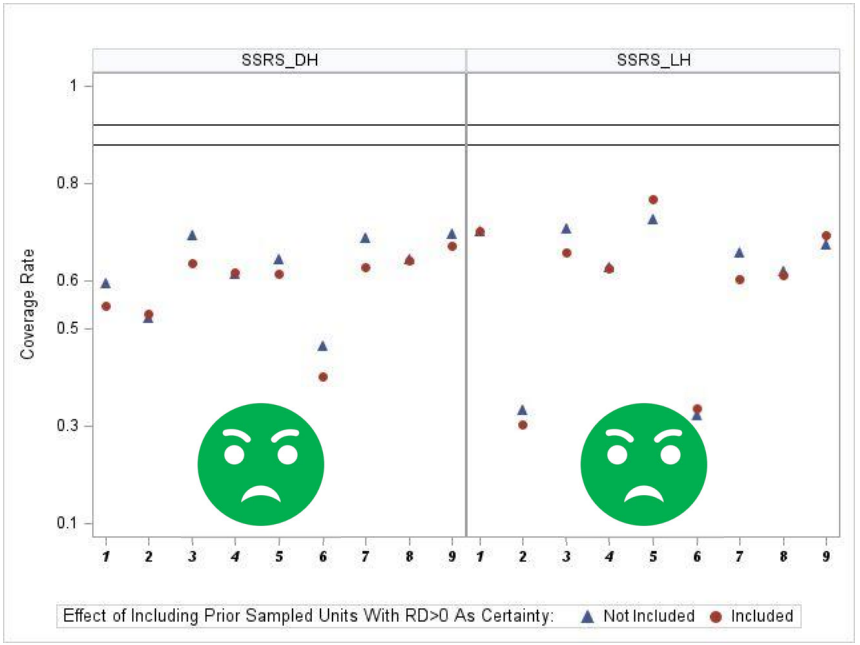
Unit Census Bureau

Coverage: One Previous Sample (Case 2)

R&D Prevalence (Proportion with R&D)



Total R&D Expenditures



Unit Census Bureau

What Have We Learned So Far…

- Evidence for
  - Using size-based stratification in sample design
  - Including certainty stratum in sample design

- Evidence against
  - Unequal probability sampling
    - Annual payroll as measure of size

United States® Census Bureau

But Still…

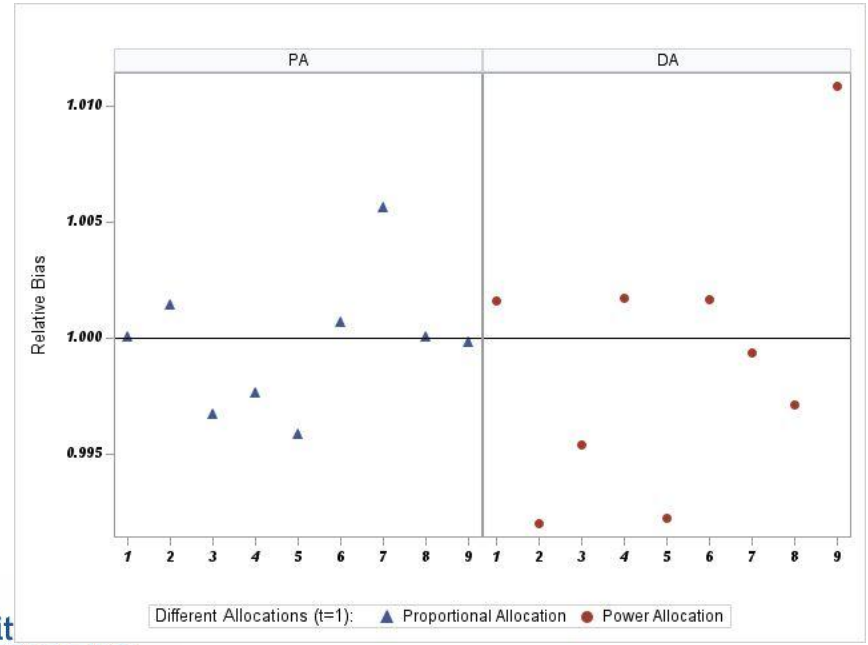
- There is severe undercoverage for total R&D expenditures
  - Best design (SSRS_LH)
  - All unit size/propensity combinations

- Could we improve results by using a different allocation scheme?
  - Power allocation
    - Allocate proportionally to the square-root of stratum-level estimate of total R&D expenditures
    - Try for case 2 (one previous sample)
  - Disproportionate allocation
    - Allocate proportionally to the square-root of the stratum level estimate of R&D prevalence
    - Not shown

United States® Census Bureau

Relative Bias: One Previous Sample (Case 2) Stratified SRS-WOR with LH Certainty Stratum

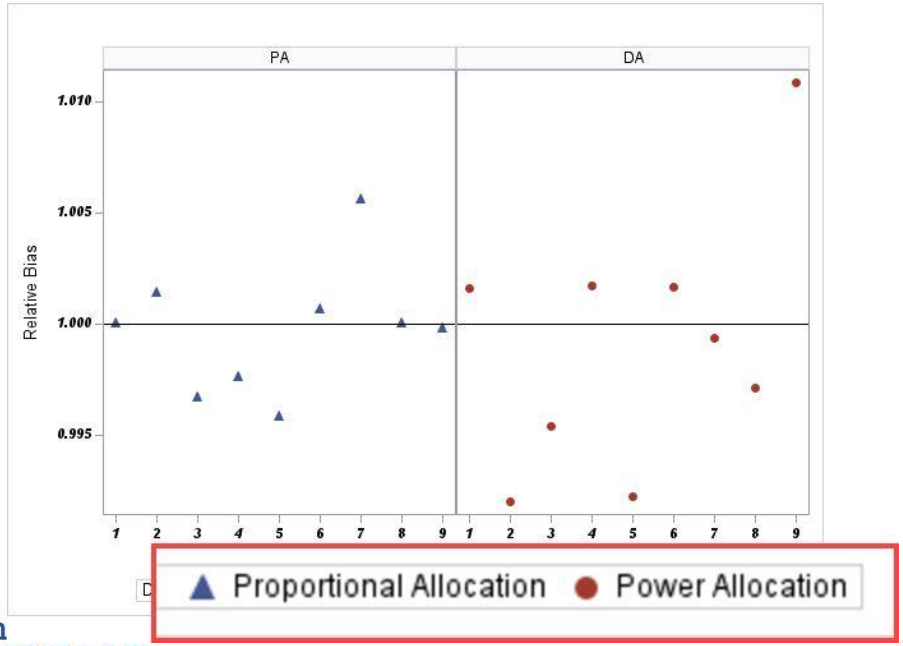
R&D Prevalence (Proportion with R&D)



Unit Census Bureau

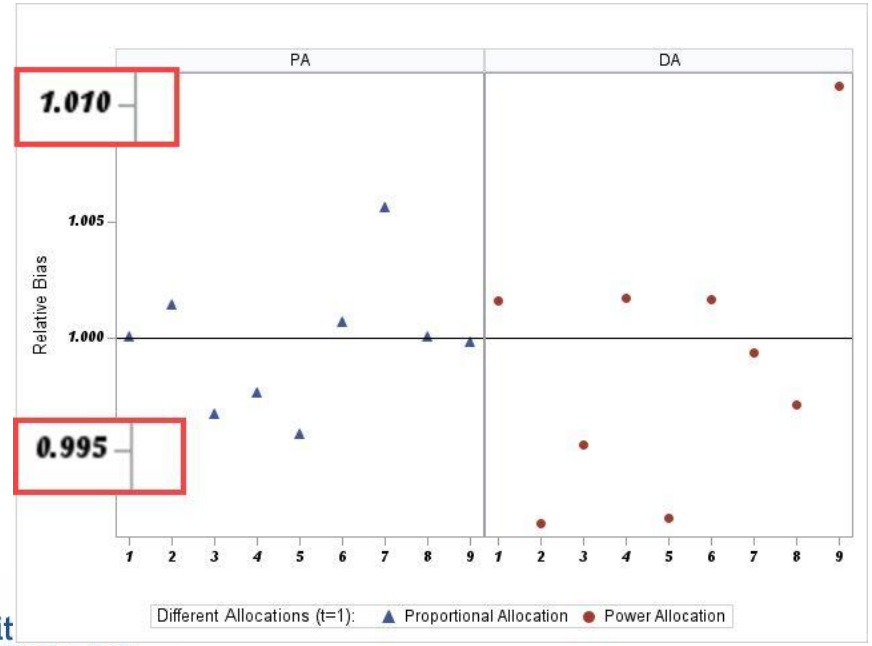
Relative Bias: One Previous Sample (Case 2) Stratified SRS-WOR with LH Certainty Stratum

R&D Prevalence (Proportion with R&D)



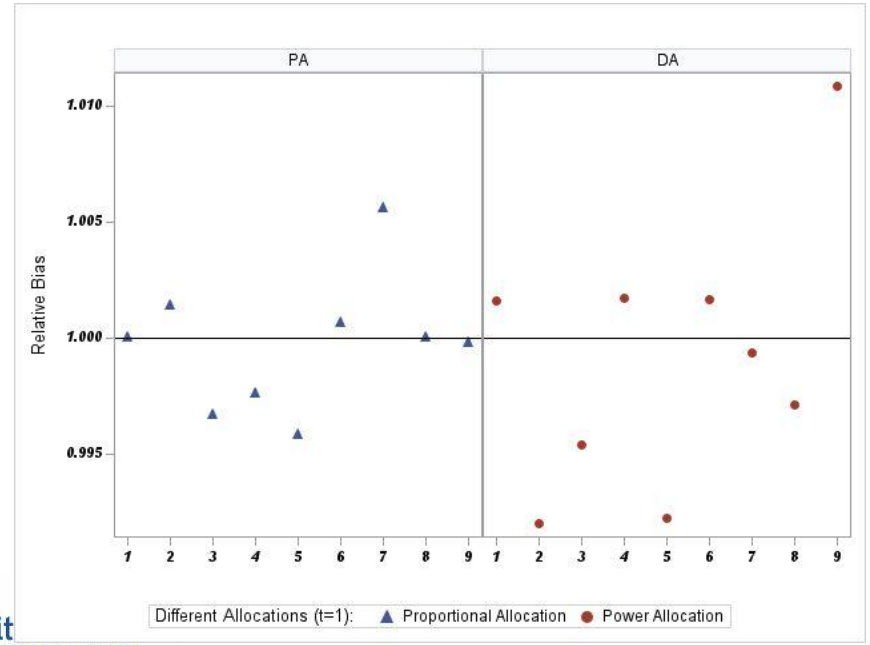
Relative Bias: One Previous Sample (Case 2) Stratified SRS-WOR with LH Certainty Stratum

R&D Prevalence (Proportion with R&D)

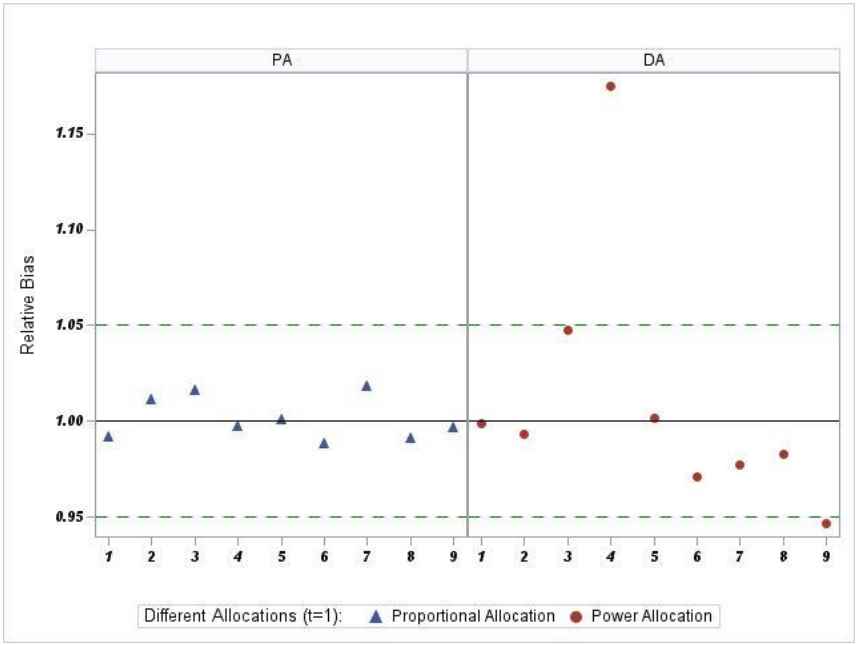


Relative Bias: One Previous Sample (Case 2) Stratified SRS-WOR with LH Certainty Stratum

R&D Prevalence (Proportion with R&D)



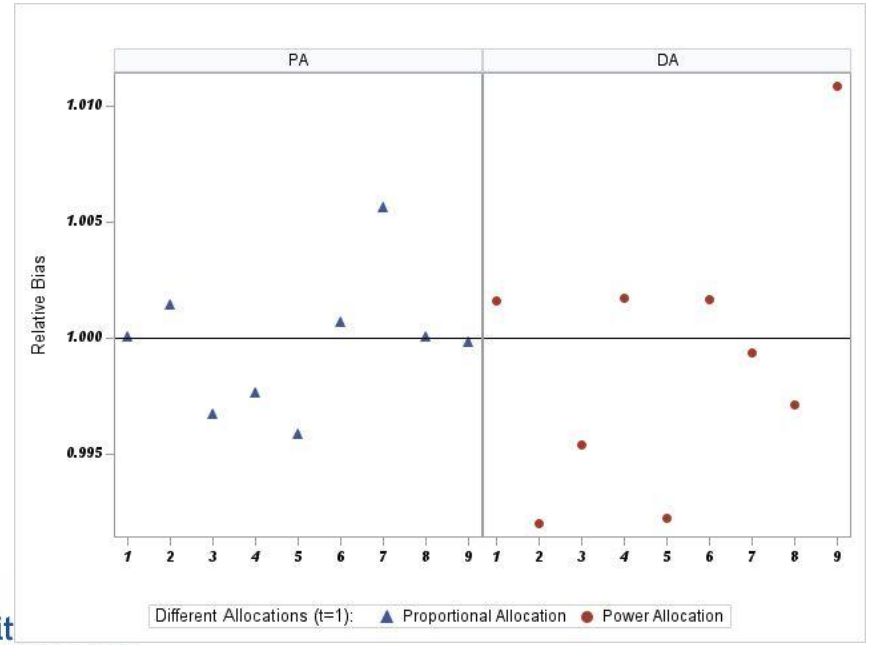
Total R&D Expenditures



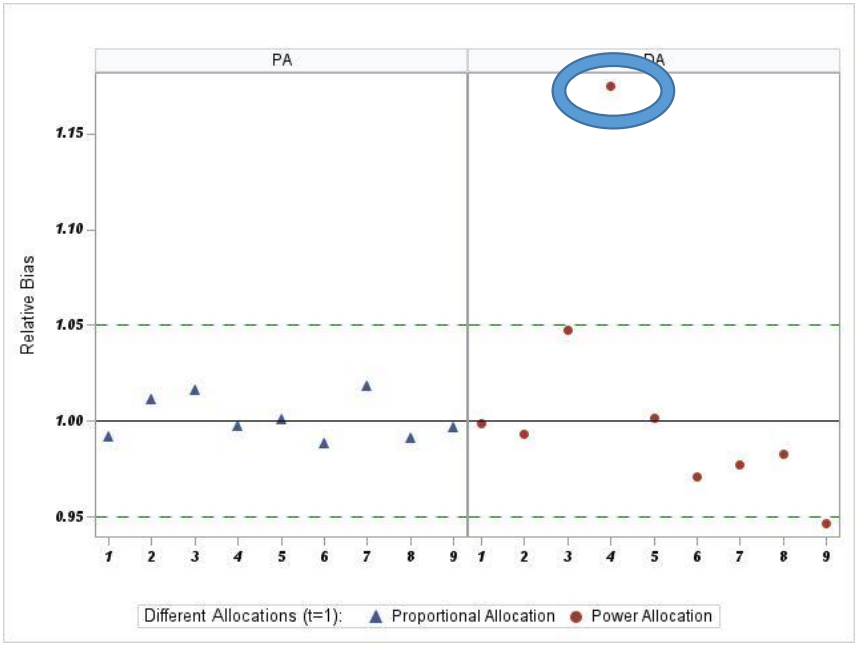
Unit Census Bureau

Relative Bias: One Previous Sample (Case 2) Stratified SRS-WOR with LH Certainty Stratum

R&D Prevalence (Proportion with R&D)



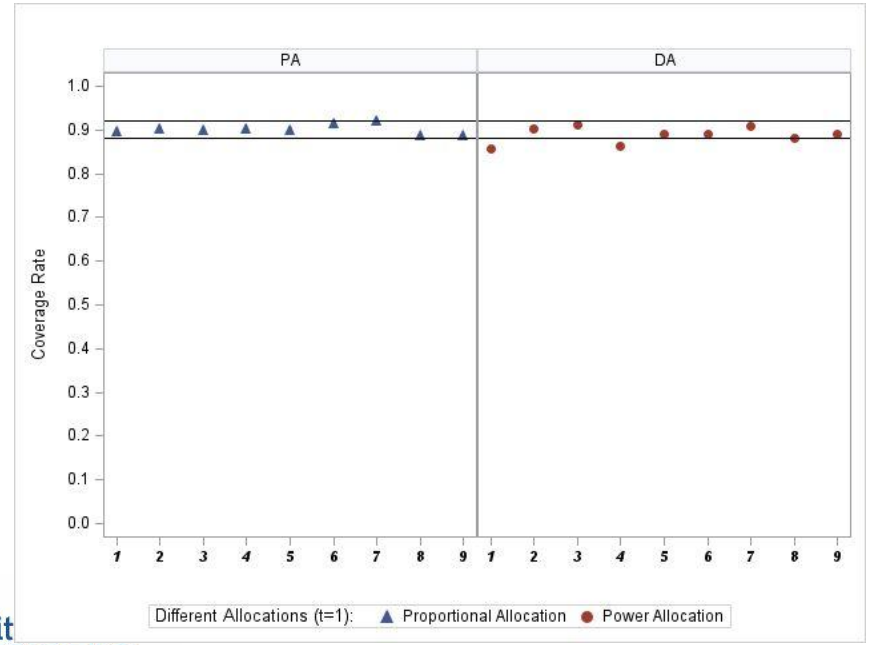
Total R&D Expenditures



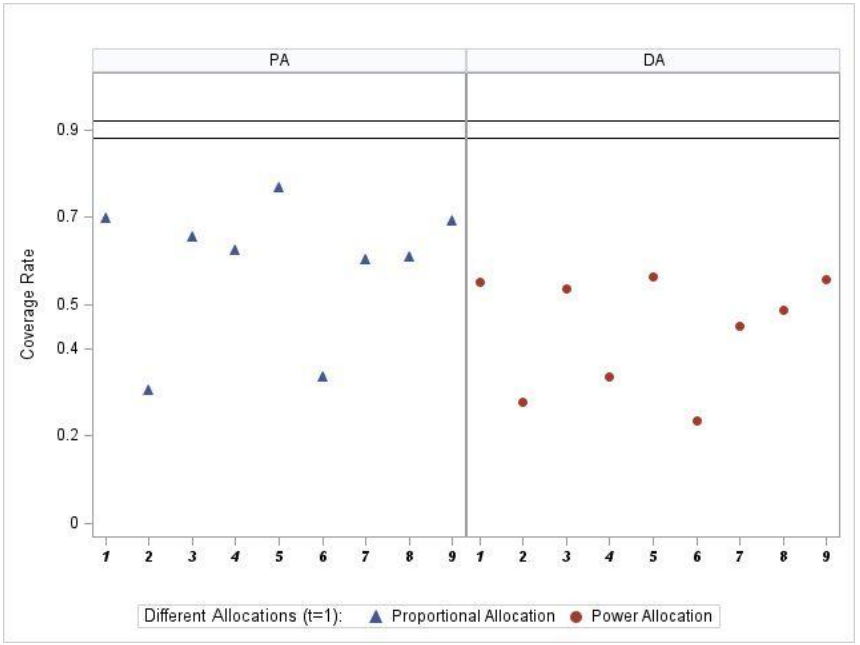
Unit Census Bureau

90% CI Coverage Rate: Case 2
Stratified SRS-WOR with LH Certainty Stratum

R&D Prevalence (Proportion with R&D)



Total R&D Expenditures



Unit Census Bureau

Let's Recap

- We proposed
  - Generate multiple partially synthetic frames (rare characteristic synthesized)
  - Draw repeated samples from each synthetic frame with one or more candidate designs
  - Assess finite sample performance for each candidate design within and between the synthetic frames

- In our case study, we found
  - Evidence for
    - Using size-based stratification in sample design
    - Including certainty stratum in sample design
  - Evidence against
    - Unequal probability sampling (annual payroll as measure of size)

United States® Census Bureau

Let's Recap

- In our case study, we did not find a sample design that balances reliability considerations for a proportion and for a total
  - But, we can try other sample designs
    - Different sampling procedures, stratifications, allocation methods

- In our case study, we found
  - Sensitive/spurious results for some sample design features
  - Consistent results for some design features (ROBUSTNESS!)
  - Deficiencies in "best sample design" that need to addressed…or acknowledged

United States® Census Bureau

A Few Final Thoughts

- We might not be especially original thinkers
  - Bayesian synthesizer ≈ superpopulation model
    - We retain the same frame, so not quite the same
    - Could expand to fully synthetic data
    - Could add more time periods

- Thanks for your attention
  - Work is still in progress
  - I welcome your suggestions, critiques, and input
  - katherine.j.thompson@census.gov

United States® Census Bureau