Using Partially Synthetic Frames To Evaluate Alternative Sample Designs For Estimating A Rare Business Characteristic

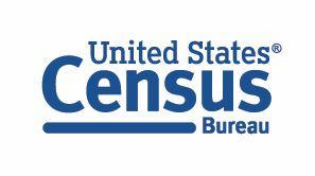
Katherine Jenny Thompson, U.S. Census Bureau
JPSM Seminar Series
October 4, 2023

Work conducted jointly with Dr. Hang J. Kim (University of Cincinnati) and Stephen Kaputa (U.S. Census Bureau)

United States® Census Bureau

Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau. 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-7504682, Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD010-021).

Note: all illustrative examples are FICTIONAL. Presented results are not.



Using Partially Synthetic Frames To Evaluate Alternative Sample Designs For Estimating A Rare Business Characteristic

United States® Census Bureau

Establishment Surveys

"measure the behavior, structure, or output of organizations rather than individuals."

The Encyclopedia of Survey Research Methods

United States® Census Bureau

Establishment Surveys Collect Data On

- **Businesses**
- Farms
- Institutions (schools, governments, prisons)

Basically, anything other than a household or a person.

United States® Census Bureau

Definitions (Economic/Business Surveys)

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

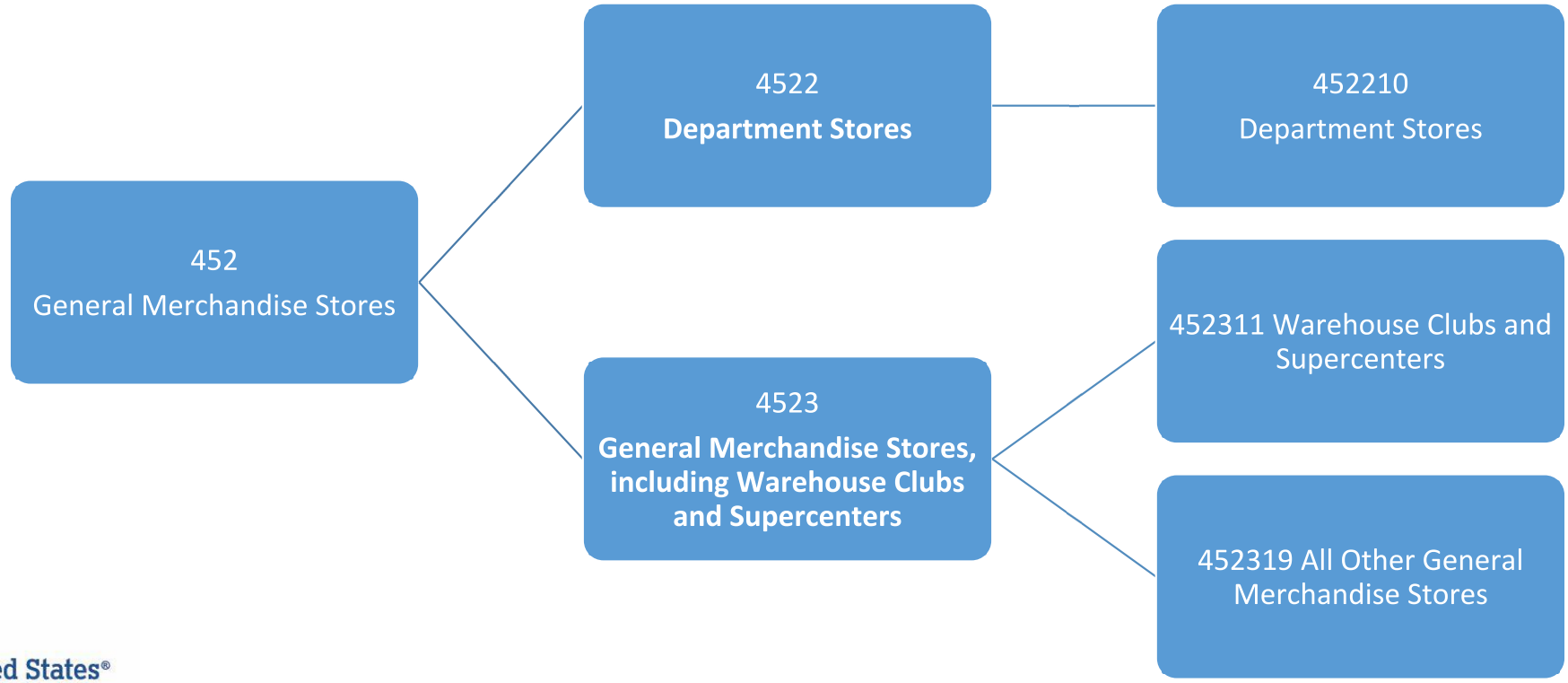
- Industry: a group of companies that are related based on their primary business activities or service (www.investopedia.com)

- 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 (44_45)



United States® Census Bureau

Company (Firm)

- Any formal business entity for profit, which may be a corporation, a partnership, association or individual proprietorship (https://dictionary.law.com/)

Company (Firm)

- Any formal business entity for profit, which may be a corporation, a partnership, association or individual proprietorship (https://dictionary.law.com/)
  - Single-unit company = individual proprietorship

Company (Firm)

- Any formal business entity for profit, which may be a corporation, a partnership, association or individual proprietorship (https://dictionary.law.com/)
  - Single-unit company = individual proprietorship

United States® Census Bureau

Company (Firm)

- Any formal business entity for profit, which may be a corporation, a partnership, association or individual proprietorship (https://dictionary.law.com/)
  - Single-unit company = individual proprietorship
  - Multi-unit company

Company (Firm)

- Any formal business entity for profit, which may be a **corporation**, a partnership, association or individual proprietorship (https://dictionary.law.com/)
  - Single-unit company = individual proprietorship
  - Multi-unit company

United States® Census Bureau

Multi-Unit Company (Firm)

Location 1: Retail Trade Sector

Location 2: Retail Trade Sector

Location 3: Retail Trade Sector

United States® Census Bureau

Multi-Unit Company (Firm)

Location 1: Retail Trade Sector

Location 2: Retail Trade Sector

Location 3: Retail Trade Sector

Location 4: Wholesale Trade Sector

United States® Census Bureau

Multi-Unit Company (Firm)

Location 1: Retail Trade Sector

Location 5: Manufacturing Sector

Location 2: Retail Trade Sector

Location 3: Retail Trade Sector

Location 4: Wholesale Trade Sector

United States® Census Bureau

Using Partially Synthetic Frames To Evaluate Alternative Sample Designs For Estimating A Rare Business Characteristic

United States® Census Bureau

Context: Finite Population Sampling

- Sampling frame
  - Complete list of eligible units
  - Contains auxiliary variables
    - Categorical
    - Continuous
- Sampling design
  - Known probability of inclusion for each unit on population
  - Predetermined (random) selection procedure
  - (Generally) Fixed sample size

United States® Census Bureau

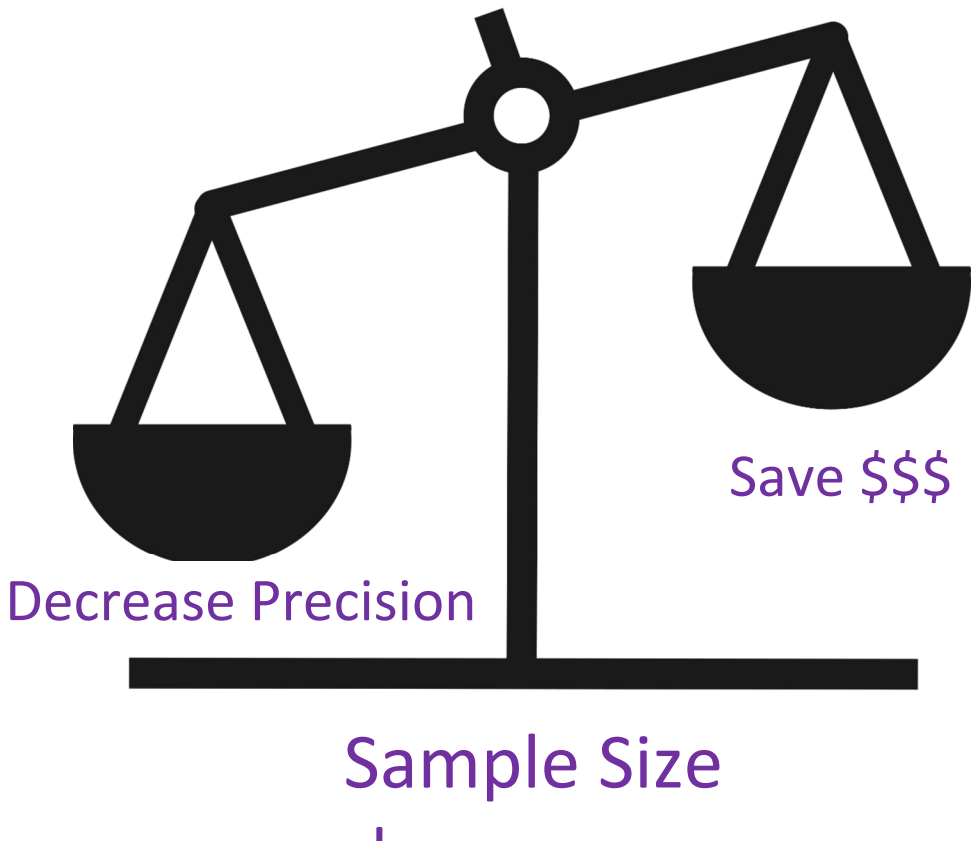
Context: Finite Population Sampling

- Budget constraints
  - FIXED sample size
- Reliability constraints

United States® Census Bureau

Context: Finite Population Sampling

- Budget constraints
  - FIXED sample size
- Reliability constraints



United States® Census Bureau

Context: Finite Population Sampling

- Budget constraints
  - FIXED sample size
- Reliability constraints

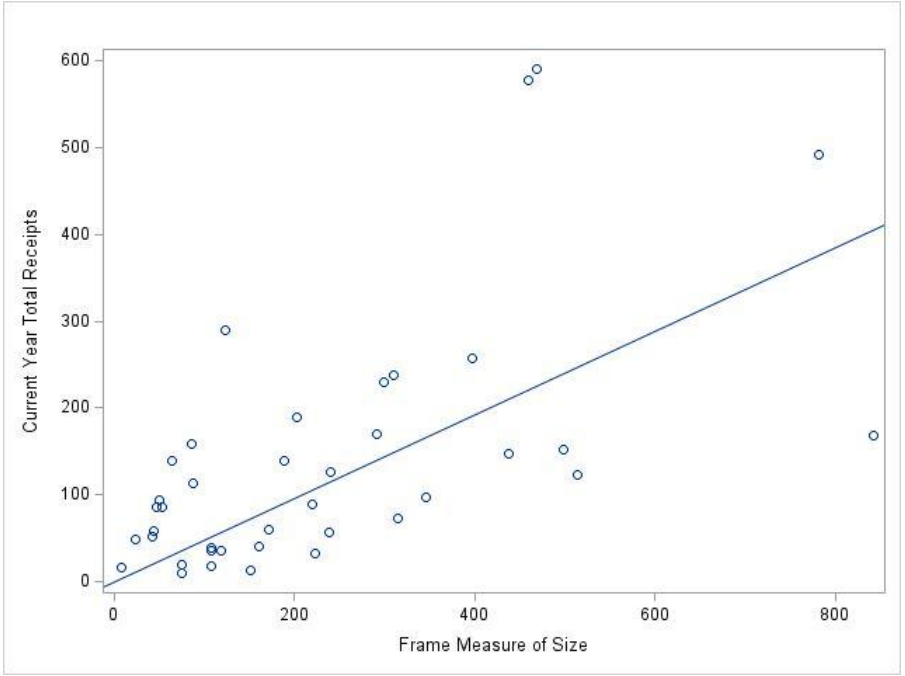
United States® Census Bureau

Mandatory Reminder – all examples are FICTIONAL

(until I get to the results)

“Typical” (Business) Survey Design Setting

- Sampling frame
  - Auxiliary variable(s)
  - Available for all units
  - Measure of size
    - Continuous
    - Positive association with survey items
- Survey items (characteristics)
  - Collected from sampled units

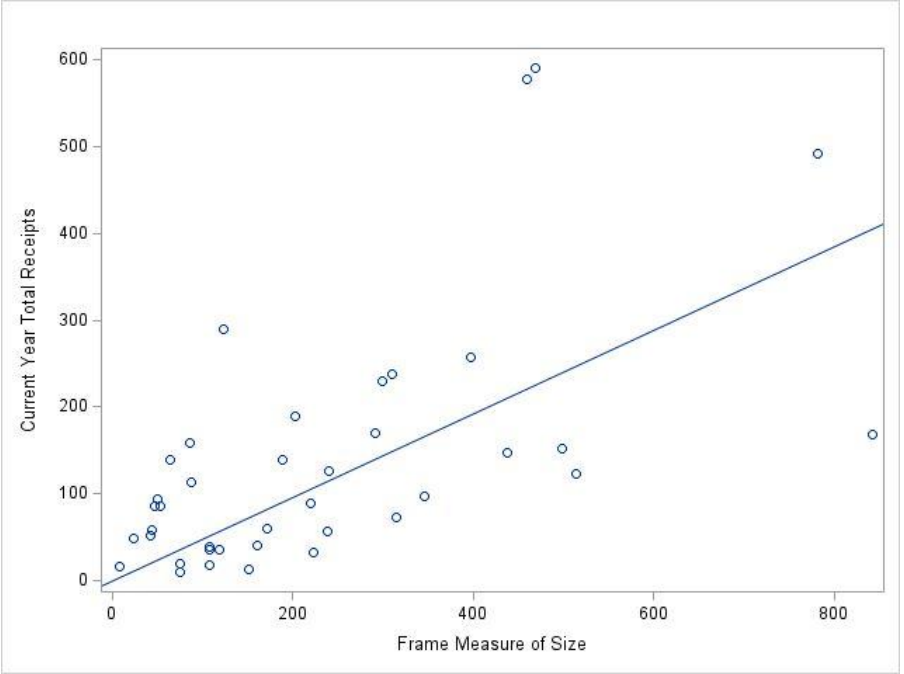


United States® Census Bureau

“Typical” (Business) Survey Design Setting

- Sample design utilizes frame measure of size

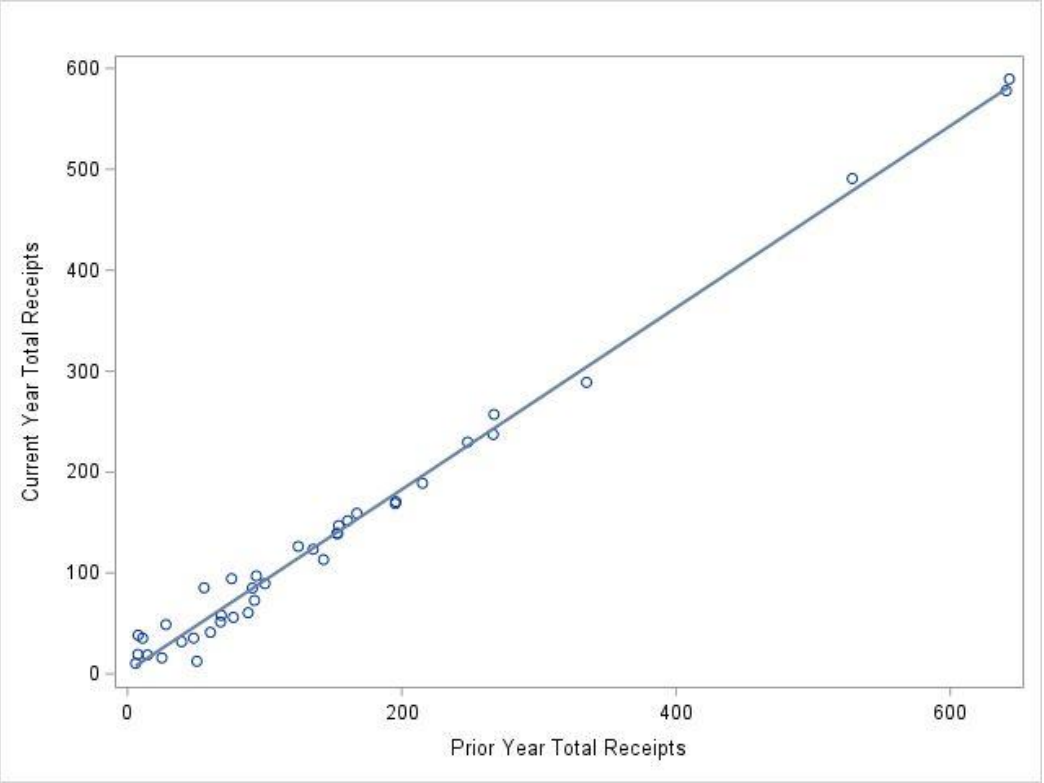
- Frame variables used to evaluate effectiveness of design
  - One or more candidate sample designs
  - Assess performance by
    - Drawing one sample from single frame
    - Drawing repeated samples from single frame



United States® Census Bureau

“Typical” Survey Data Assumption

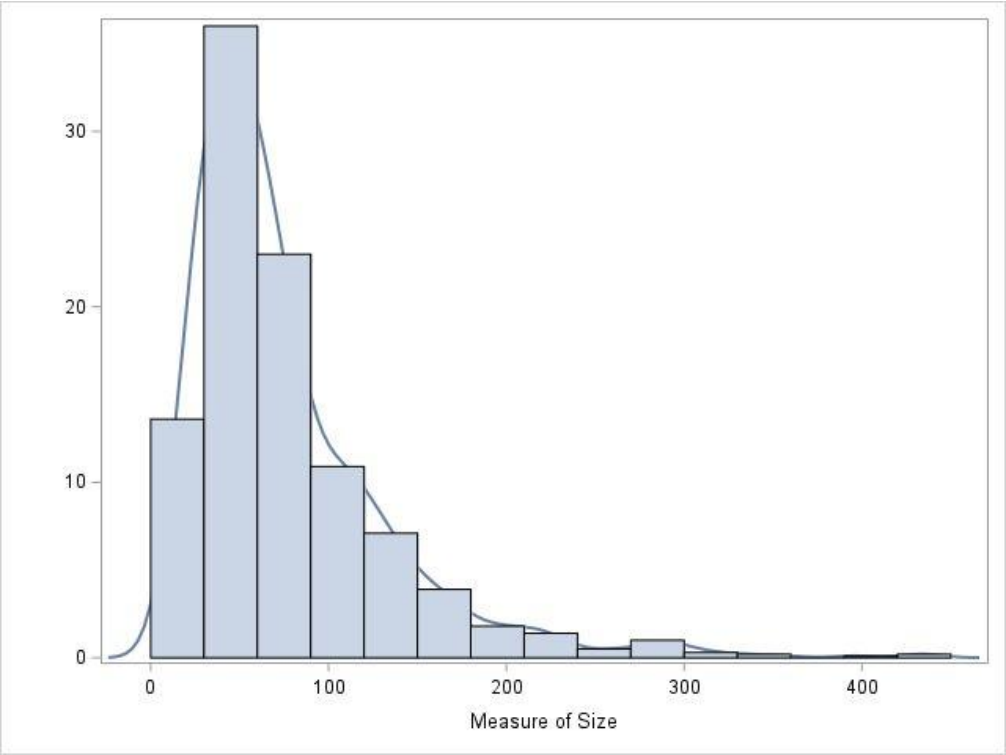
- Strong positive association between prior reported value and current reported value
- Discrete variables
  - Prior <u>state</u> is a predictor of current state
- Continuous variables (pictured)
  - Prior <u>value</u> is a predictor of current value



United States® Census Bureau

Establishment (Business) Sample Survey Designs

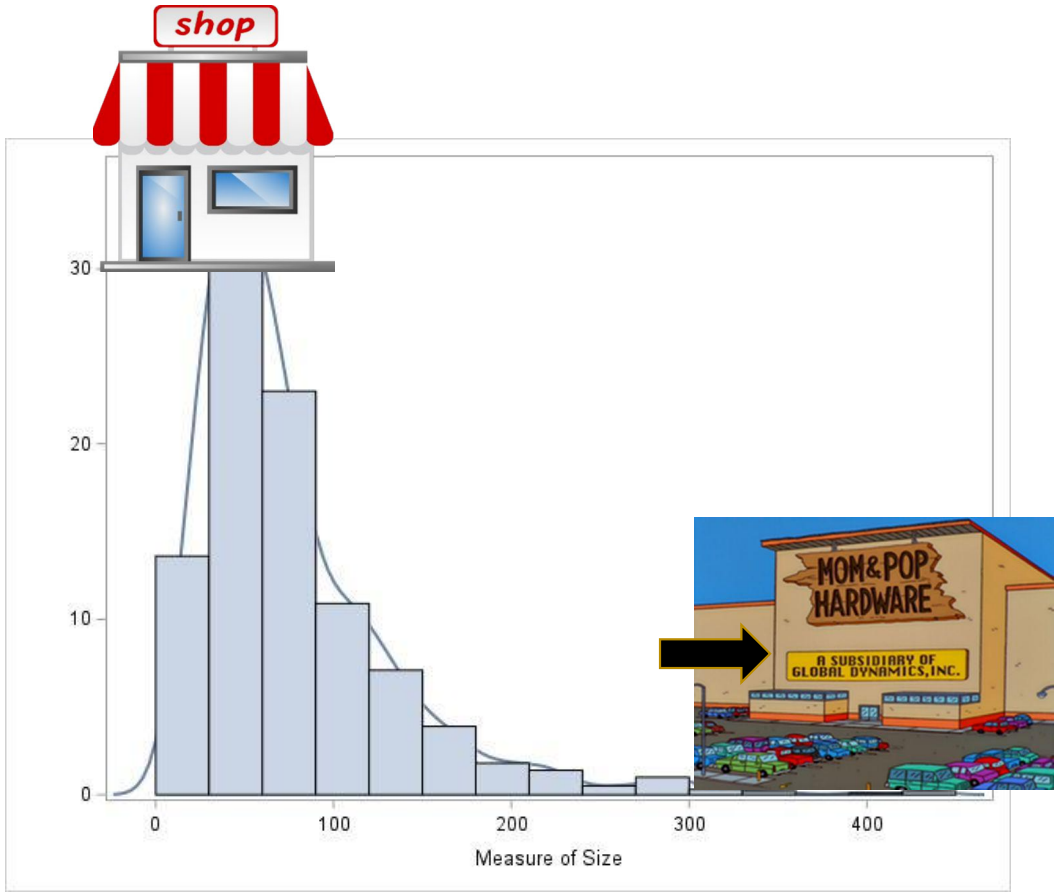
- Populations are skewed!
  - Small number of large companies
  - Majority small companies



United States® Census Bureau

Establishment (Business) Sample Survey Designs

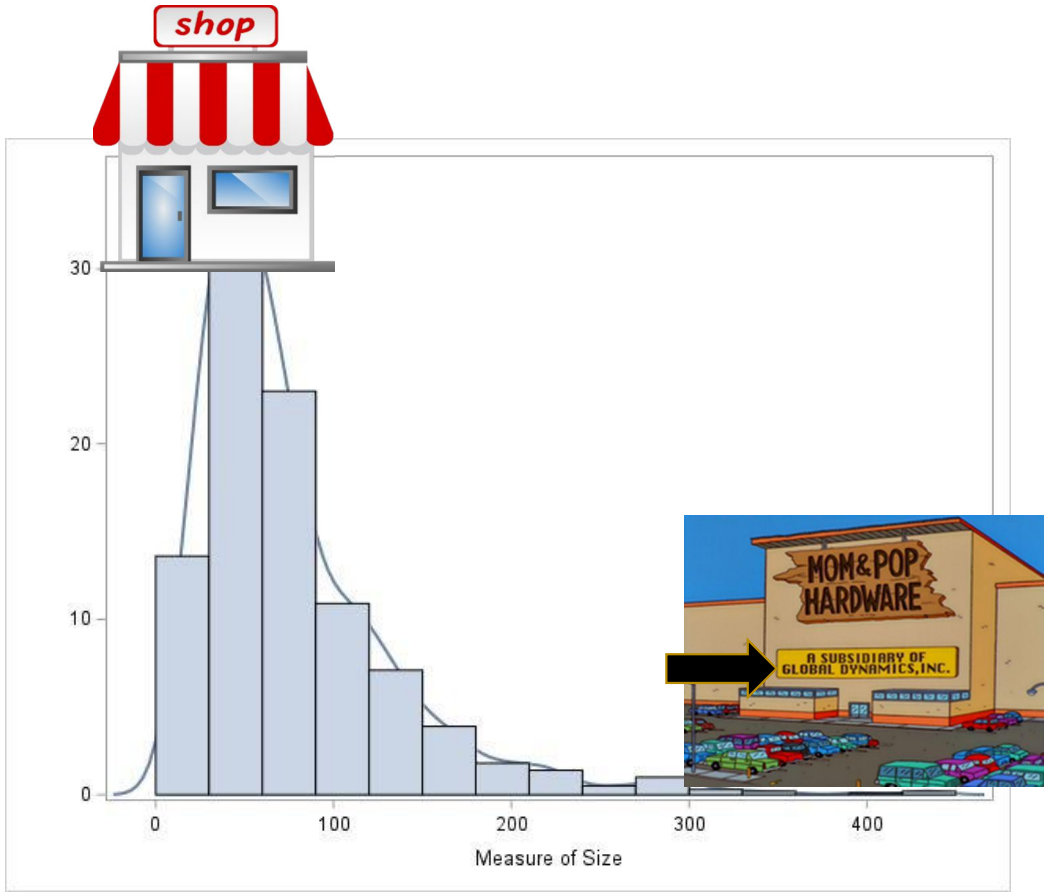
- Populations are skewed!
  - Small number of large companies
  - Majority small companies



United States® Census Bureau

Establishment (Business) Sample Survey Designs

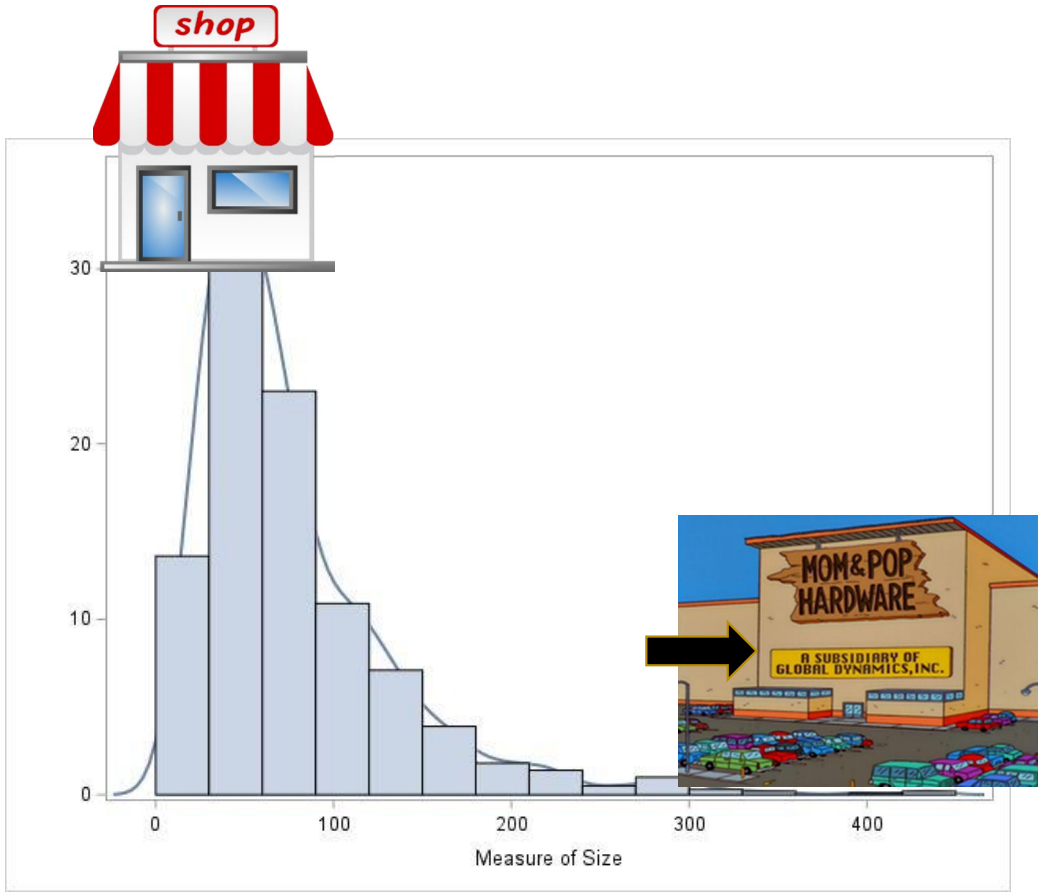
- Populations are skewed!
  - Small number of large companies
  - Majority small companies
- Publish TOTALS
  - And ratios of totals



United States® Census Bureau

Establishment (Business) Sample Survey Designs

- Populations are skewed!
  - Small number of large companies
  - Majority small companies
- Publish TOTALS
  - And ratios of totals
- **Sample design must account for skewed population**



United States® Census Bureau

Using Partially Synthetic Frames To Evaluate Alternative Sample Designs For Estimating A Rare Business Characteristic

United States® Census Bureau

Our Setting

- The frame provides a complete list of eligible units
- The auxiliary variables on the frame are ***weakly*** related to the characteristic(s) of interest
- The characteristic(s) of interest is (are) ***rare***

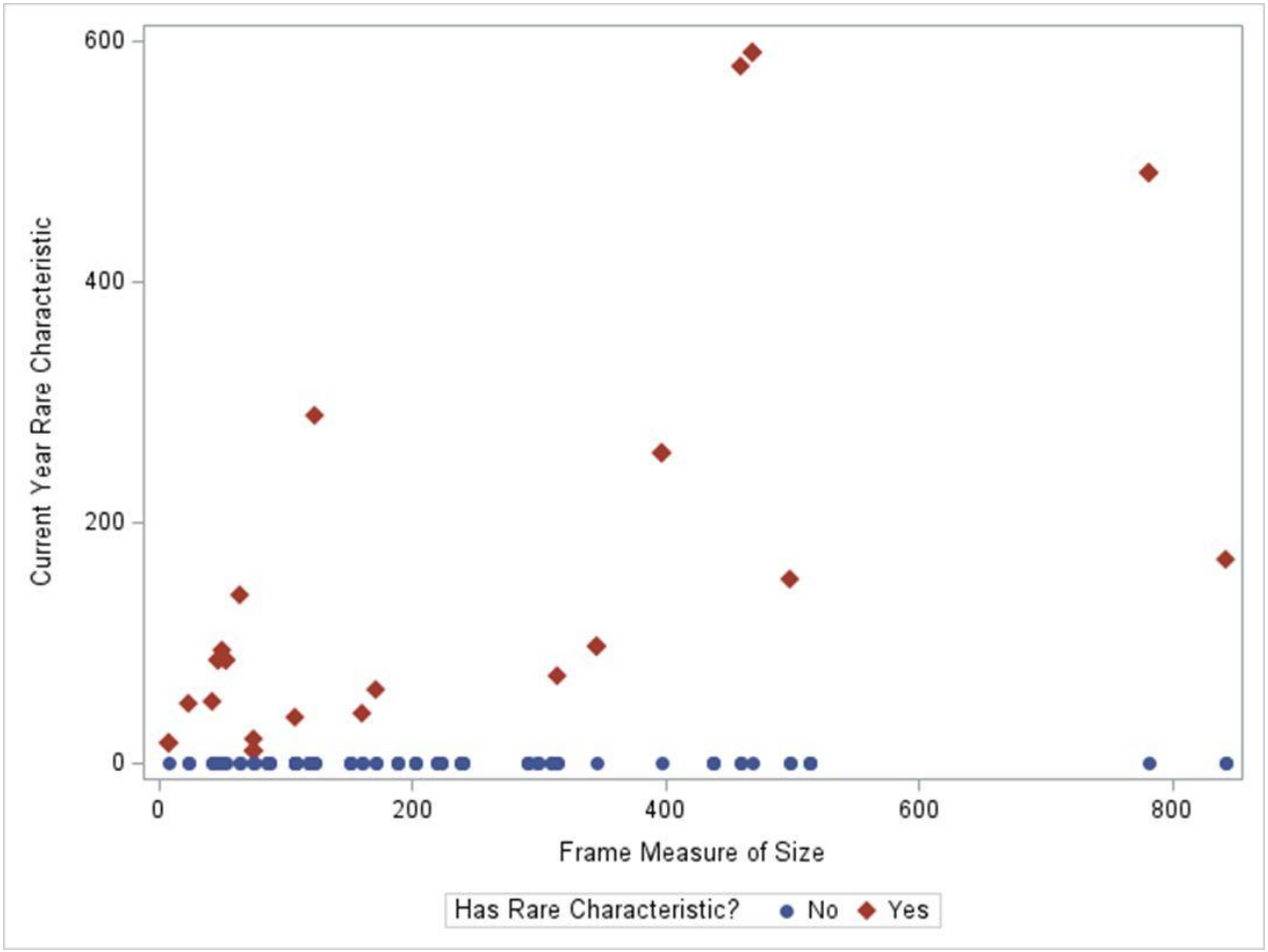
United States® Census Bureau

Network Sampling is NOT an Option

- Budget constraints (FIXED sample size)
- Reliability constraints
- Spatial clustering assumption could be tenuous
- Unreasonable to expect that a business would provide reliable information about its potential competitor(s)

United States® Census Bureau

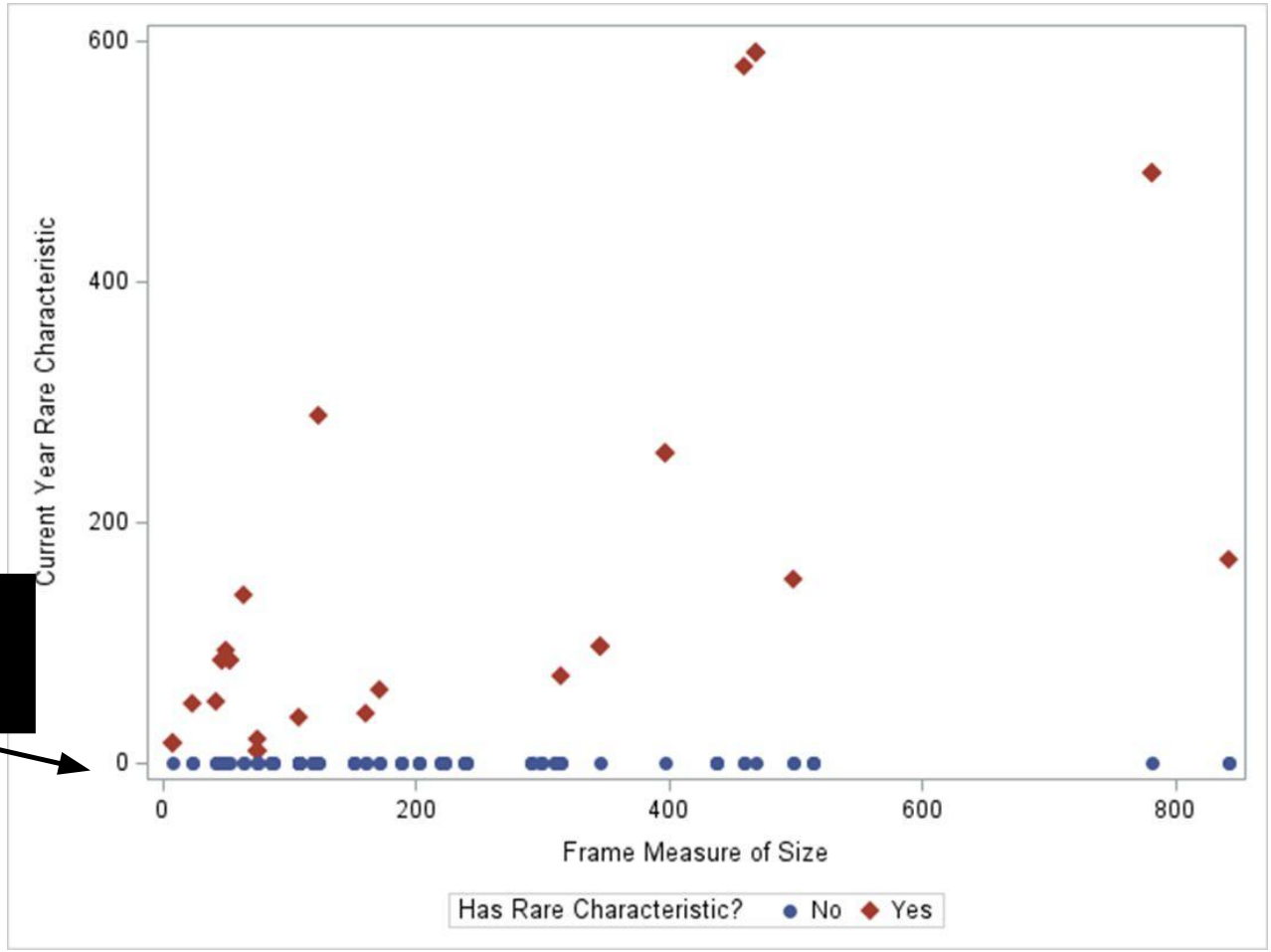
An Illustration …



United States® Census Bureau

An Illustration…

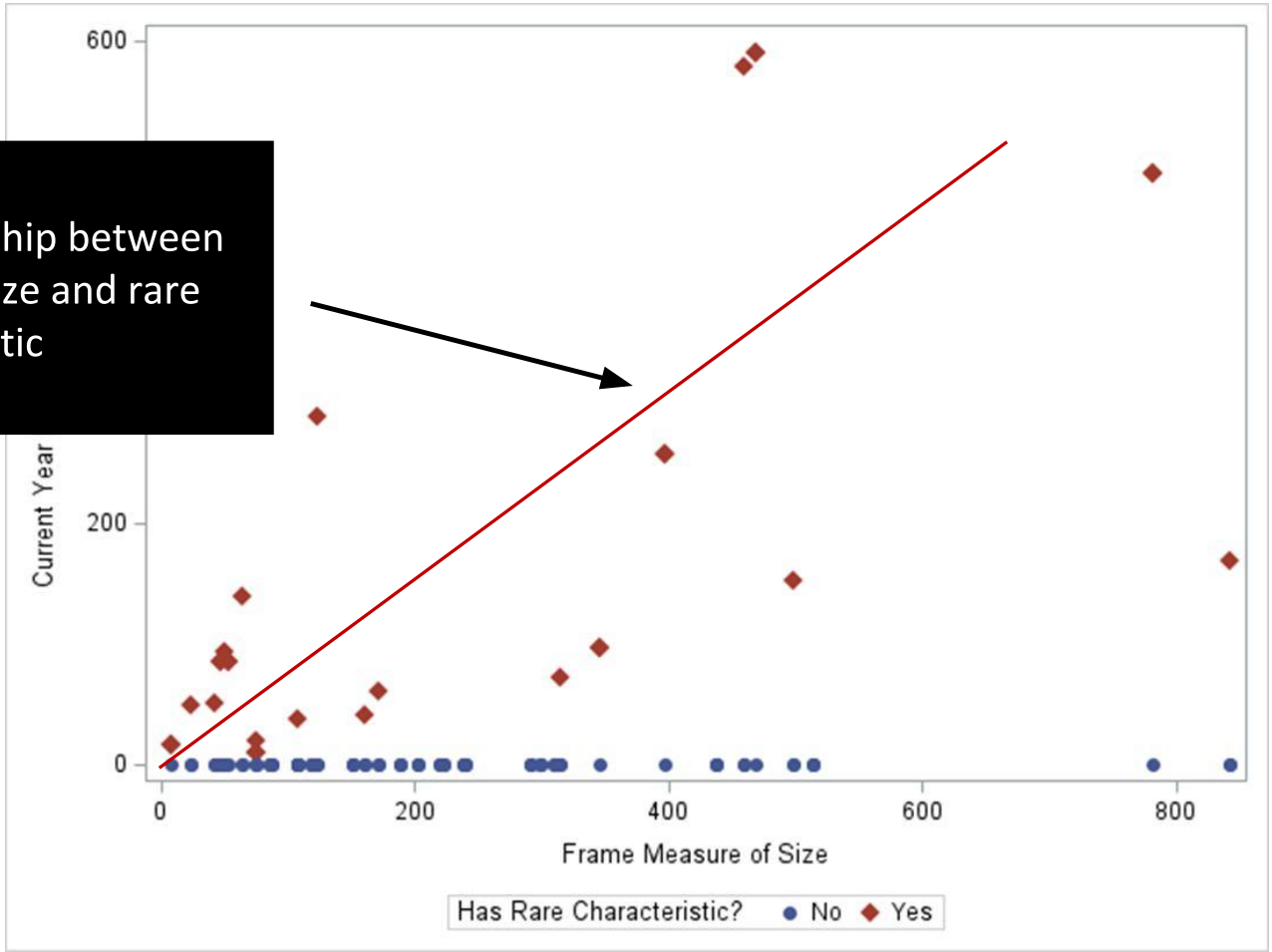
Characteristic is <u>not</u> present in most observations



United States® Census Bureau

An Illustration…

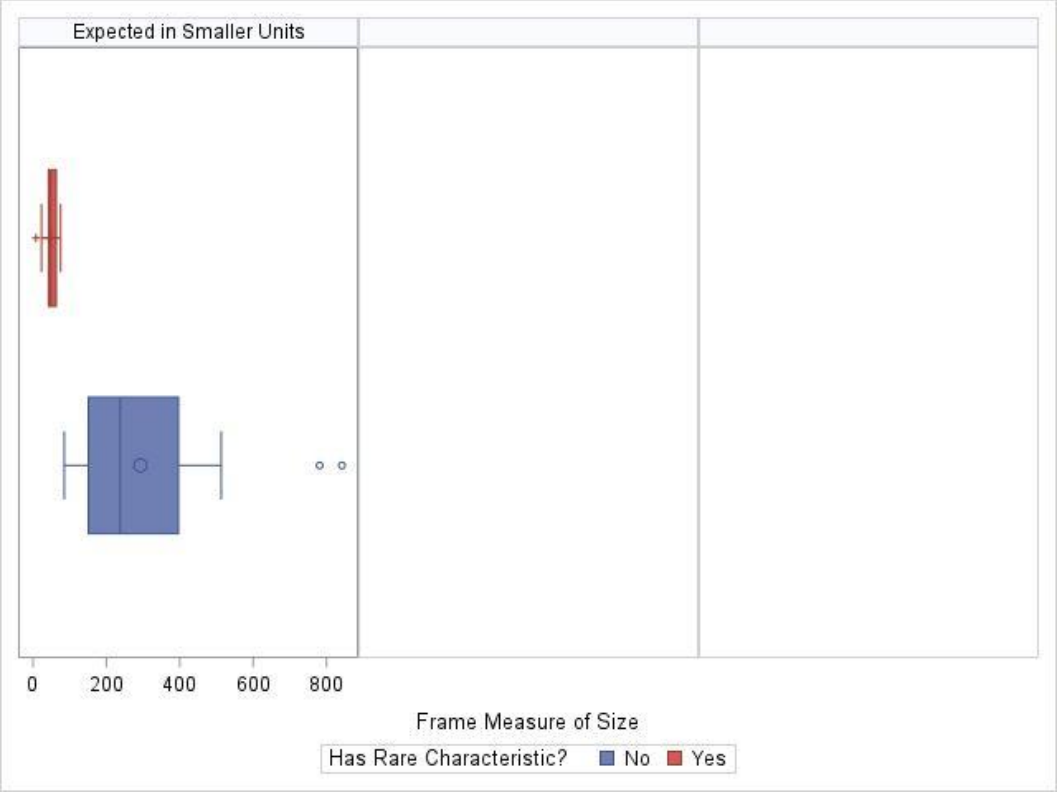
Weak linear relationship between frame measure of size and rare characteristic



United States® Census Bureau

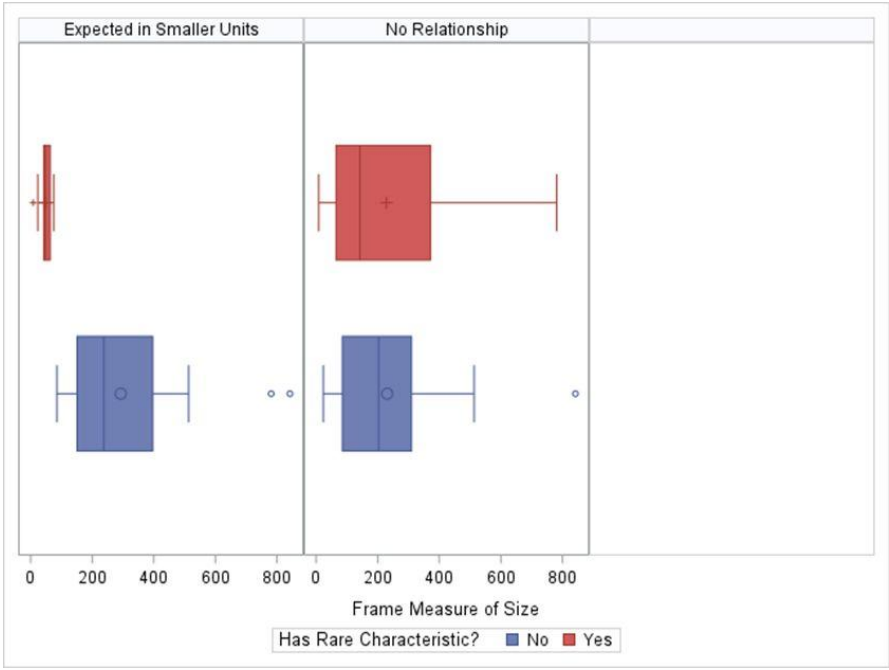
Association between the frame measure of size (MOS) and the rare characteristic could be

- Expected to appear primarily in smaller units



Association between the frame measure of size (MOS) and the rare characteristic could be

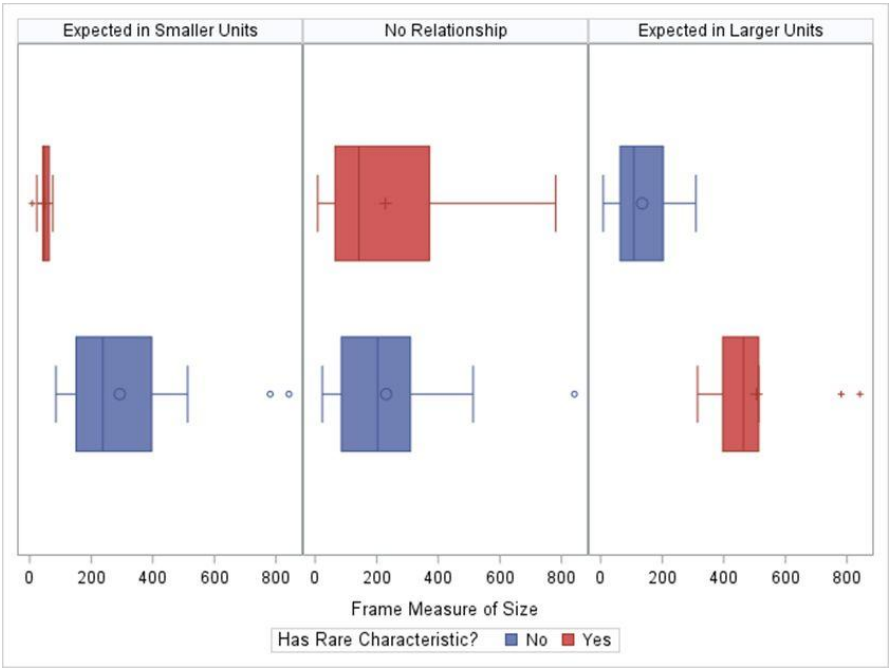
- Expected to appear primarily in smaller units
- Essentially nonexistent



United States® Census Bureau

Association between the frame measure of size (MOS) and the rare characteristic could be

- Expected to appear primarily in smaller units
- Essentially nonexistent
- Expected to appear primarily in larger units



United States® Census Bureau

Using Partially Synthetic Frames To Evaluate Alternative Sample Designs For Estimating A Rare Business Characteristic

United States® Census Bureau

Synthetic Data

- General procedure
  - Develop model using the original data
  - Release repeated draws from the model (posterior predictive distribution)
- Two flavors
  - Fully synthetic data (all variables synthesized)
  - Partially synthetic data (some variables synthesized)
- Generally used in data confidentiality applications

United States® Census Bureau

Partially Synthetic Frames

ID									
00001									
000002									
000003									
000004									
000005									
000006									
000007									
000008									
...									
001214									

United States® Census Bureau

FRAME VARIABLES

ID	NAICS	State	X (MOS)						
000001	452311	MO	39						
000002	452311	TN	11						
000003	452311	MO	18						
000004	452311	TN	59						
000005	452311	MO	71						
000006	452319	MO	21						
000007	452319	MO	42						
000008	452319	TN	91						
...									
001214	452311	MO	3						

United States® Census Bureau

SAMPLE DATA VARIABLES

ID	NAICS	State	X (MOS)	Sample Indicator	Y (Survey)	Design Weight			
000001	452311	MO	39	1	28	1.5			
000002	452311	TN	11	1	9	30			
000003	452311	MO	18	1	11	30			
000004	452311	TN	59	1	46	1			
000005	452311	MO	71	0					
000006	452319	MO	21	1	18	2			
000007	452319	MO	42	0					
000008	452319	TN	91	0					
...									
001214	452311	MO	3	1	2	30			

United States® Census Bureau

FRAME VARIABLES

SAMPLE DATA VARIABLES

ID	NAICS	State	X (MOS)	Sample Indicator	Y (Survey)	Design Weight			
000001	452311	MO	39	1	28	1.5			
000002	452311	TN	11	1	9	30			
000003	452311	MO	18	1	11	30			
000004	452311	TN	59	1	46	1			
000005	452311	MO	71	0					
000006	452319	MO	21	1	18	2			
000007	452319	MO	42	0					
000008	452319	TN	91	0					
...									
001214	452311	MO	3	1	2	30			

United States® Census Bureau

FRAME VARIABLES

SAMPLE DATA VARIABLES

ID	NAICS	State	X (MOS)	Sample Indicator	Y (Survey)	Design Weight			
000001	452311	MO	39	1	28	1.5			
000002	452311	TN	11	1	9	30			
000003	452311	MO	18	1	11	30			
000004	452311	TN	59	1	46	1			
000005	452311	MO	71	0					
000006	452319	MO	21	1	18	2			
000007	452319	MO	42	0					
000008	452319	TN	91	0					
...									
001214	452311	MO	3	1	2	30			

Fit model using frame data (NAICS, State, X) to predict survey item (Y)

United States® Census Bureau

SYNTHETIC DATA

ID	NAICS	State	X (MOS)	Sample Indicator	Y (Survey)	Design Weight	Y(1)	Y(2)	…	Y(M)
000001	452311	MO	39	1	28	1.5	34	37		33
000002	452311	TN	11	1	9	30	14	7		13
000003	452311	MO	18	1	11	30	17	10		18
000004	452311	TN	59	1	46	1	43	46		25
000005	452311	MO	71	0			61	23		52
000006	452319	MO	21	1	18	2	14	19		28
000007	452319	MO	42	0			57	51		42
000008	452319	TN	91	0			89	65		80
…										
001214	452311	MO	3	1	2	30	4	3		2

United States® Census Bureau

SYNTHETIC DATA

ID	NAICS	State	X (MOS)	Sample Indicator	Y (Survey)	Design Weight	Y(1)	Y(2)	…	Y(M)
000001	452311	MO	39				34	37		33
000002	452311	TN	11				14	7		13
000003	452311	MO	18				17	10		18
000004	452311	TN	59				43	46		25
000005	452311	MO	71				61	23		52
000006	452319	MO	21				14	19		28
000007	452319	MO	42				57	51		42
000008	452319	TN	91				89	65		80
…										
001214	452311	MO	3	1	2	30	4	3		2

M sets of (independent) predicted values of item *Y* per unit on frame

United States® Census Bureau

Using Partially Synthetic Frames To Evaluate Alternative Sample Designs For Estimating A Rare Business Characteristic

United States® Census Bureau

Our Approach

- Generate multiple partially synthetic frames
- Draw repeated samples from each synthetic frame
  - Consider one or more candidate designs
- Assess finite sample performance for each candidate design within and between the synthetic frames

United States® Census Bureau

Considerations

- Strength of association between frame variables and outcome variables(s)

- Available data for modeling (besides frame variables)
  - Historic data
  - Other source (3rd party) data

- Survey collection
  - Cross sectional
  - Longitudinal

United States® Census Bureau

Case Study: Business Enterprise ***Research and Development (R&D)*** Survey (BERD)

United States® Census Bureau

Case Study

- Business Enterprise ***Research and Development (R&D)*** Survey (BERD)
  - Annual Survey
  - Conducted by the U.S. Census Bureau in partnership with the National Science Foundation's National Center for Science & Engineering Statistics

- Key estimate is *total R&D expenditures*
  - Also estimates *R&D prevalence*

- Five study industries
  - Varying proportions of companies with characteristic

United States® Census Bureau

Case Study Data

Frame data

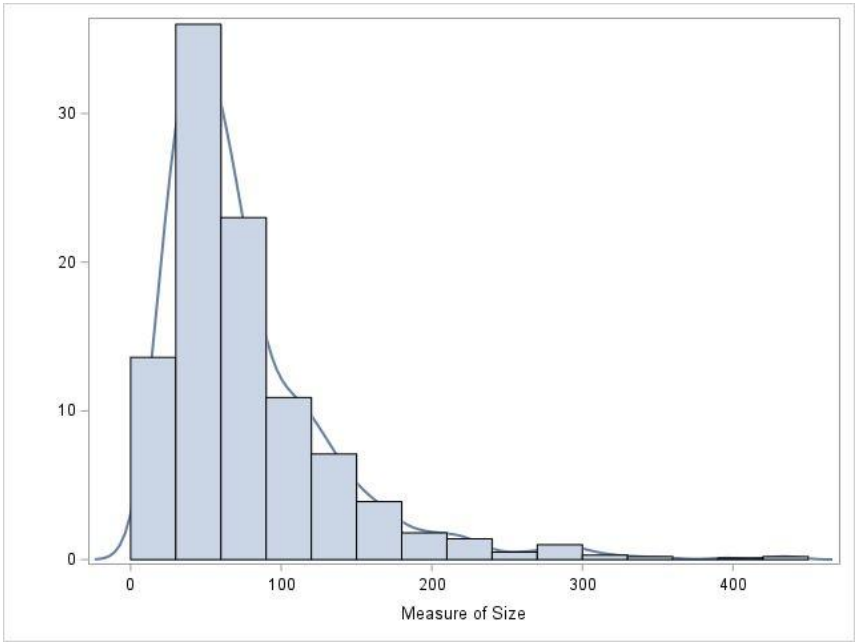
- 2018 and 2019
- Industry code (NAICS)
- Annual Payroll
  - Positively skewed within industry
  - Used as **Measure of Size (MOS)**
- Total Number of Employees
  - Positively skewed within industry
  - Used for evaluation (traditionally)

United States® Census Bureau

Case Study Data

Frame data

- 2018 and 2019
- Industry code (NAICS)
- Annual Payroll
  - Positively skewed within industry
  - Used as **Measure of Size (MOS)**
- Total Number of Employees
  - Positively skewed within industry
  - Used for evaluation (traditionally)



Illustration

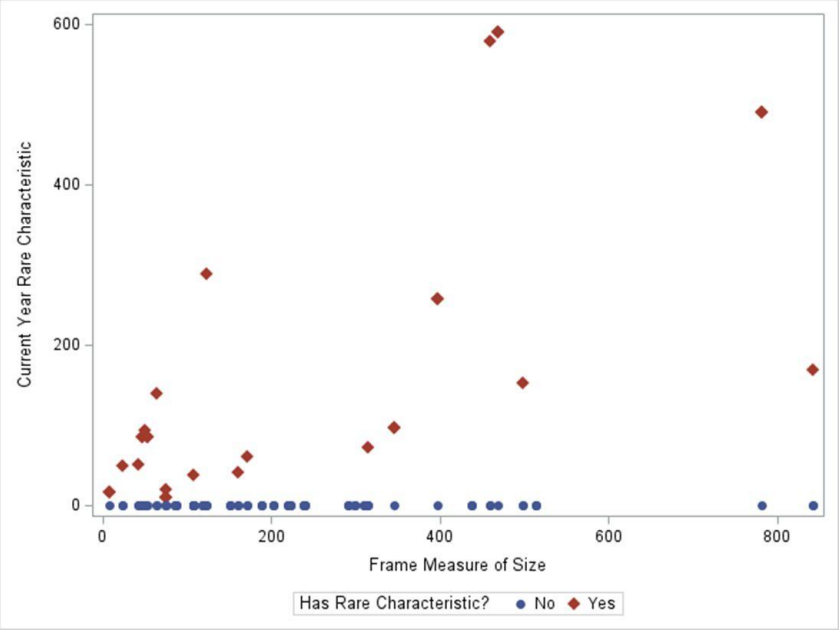
Case Study Data

Sample data

- 2018 and 2019
- Independent samples
  - High overlap of “certainty” units (sampled with probability = 1)
  - Low overlap otherwise
- Total R&D expenditures (respondents)
- Nonresponse adjusted sampling weight

United States® Census Bureau

Case Study Data



Illustration

Sample data

- 2018 and 2019
- Independent samples
  - High overlap of “certainty” units (sampled with probability = 1)
  - Low overlap otherwise
- Total R&D expenditures (respondents)
- Nonresponse adjusted sampling weight

Case Study Data

Frame data

- 2018 and 2019
- Industry code (NAICS)
- Annual Payroll
  - Positively skewed within industry
  - Used as **Measure of Size (MOS)**
- Total Number of Employees
  - Positively skewed within industry
  - Used for evaluation (traditionally)

Sample data

- 2018 and 2019
- Independent samples
  - High overlap of “certainty” units (sampled with probability = 1)
  - Low overlap otherwise
- Total R&D expenditures (respondents)
- Nonresponse adjusted sampling weight

United States® Census Bureau

Exploratory Data Analysis

- Probability of company reporting any R&D expenditures
  - Strongly associated with prior reported R&D activity
  - Not associated with unit size (Annual Payroll)
  - Not associated with auxiliary frame variable (Total Employment)

- Value of total R&D expenditures
  - Strong association with prior R&D expenditures
  - Weak association with unit size (Annual Payroll)
  - Very weak association with auxiliary frame variable (Total Employment)

United States® Census Bureau

Synthetic Data Modeling Procedure

- Merge 2018 and 2019 frames

- Create synthetic values of R&D expenditures for each company
  - Bayesian framework (Rstan)
    - Incorporate survey design data/features into models
  - Two-step process
    - Model R&D propensity
    - Model R&D expenditures value, given modeled R&D propensity
      - Case 1 (t = 0)*:* No historic data (new survey)
      - Case 2 (t = 1): One prior (independent) sample

United States® Census Bureau

Modeling R&D Propensity (0/1)

	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	Yes		
	No		

- First Collection (*t=0*)
  - No historic data
  - NO relationship between characteristic and company size
- "Weighted" coin toss

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	
First Collection (t = 0)	Yes	Sample data from 2018 collection
	No	

- First Collection (*t=0*)
  - No historic data
  - NO relationship between characteristic and company size
- "Weighted" coin toss
  - π_0=Prob(YES) at *t = 0*

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	
First Collection (t = 0)	Yes	Sample data from 2018 collection
	No	

Fit intercept-only logistic regression model

$$log\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha_0^{P0}$$

Pseudo-likelihood uses 2018 survey weights

- First Collection (*t=0*)
  - No historic data
  - NO relationship between characteristic and company size
- "Weighted" coin toss
  - π_0=Prob(YES) at *t = 0*

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	
First Collection (t = 0)	Yes	Sample data from 2018 collection
	No	

- First Collection ($t=0$)
  - No historic data
  - NO relationship between characteristic and company size
- "Weighted" coin toss
  - π_0=Prob(YES) at $t = 0$

Draw $\tilde{u}_{i0r} \sim Bernouilli\left(\frac{1}{1+e^{-\tilde{\alpha}_{0r}^{P0}}}\right)$

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	
First Collection (t = 0)	Yes	Sample data from 2018 collection
	No	

- First Collection (*t=0*)
  - No historic data
  - NO relationship between characteristic and company size
- "Weighted" coin toss
  - π_0=Prob(YES) at *t = 0*

Draw $\tilde{u}_{i0r} \sim Bernouilli\left(\frac{1}{1+e^{-\tilde{\alpha}_{0r}^{P0}}}\right)$

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	Yes		
	No		

- Second Collection (*t=1*)
  - Linked sample data
  - Evidence of relationship between prior and current R&D status
- Conditional propensity

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	Yes		
	No		

- Second Collection ($t=1$)
  - Linked sample data
  - Evidence of relationship between prior and current R&D status
- Conditional propensity
  - π_1=Prob(YES) at $t = 1$

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	Second Collection (t =1)	
		Yes	No
First Collection (t = 0)	Yes		
	No		

- Second Collection ($t=1$)
  - Linked sample data
  - Evidence of relationship between prior and current R&D status
- Conditional propensity
  - π_1=Prob(YES) at $t = 1$

Fit intercept-only logistic regression model

$$log\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha_0^{P11} + \alpha_0^{P01}$$

Pseudo-likelihood uses (2018 *x* 2019) survey weights

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	Yes		
	No		

- Second Collection (t=1)
  - Linked sample data
  - Evidence of relationship between prior and current R&D status
- Conditional propensity
  - π_1=Prob(YES) at t = 1

Additional constraint in model

$$\pi_1 = \frac{N\tilde{\pi}_0\pi_{11} + N(1-\tilde{\pi}_0)\pi_{01}}{N} \sim \text{Normal}(\hat{\pi}_1, \hat{\sigma}^2_{\hat{\pi}_1})$$

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	Second Collection (t =1)	
First Collection (t=0)		Yes	No
	Yes		
	No		

$\tilde{u}_{i0r}$ = 1

- Second Collection (t=1)
  - Linked sample data
  - Evidence of relationship between prior and current R&D status
- Conditional propensity
  - π_1=Prob(YES) at t = 1

Draw $\tilde{u}_{i1r} \sim Bernouilli\left(\frac{1}{1+e^{-\tilde{\alpha}_{0r}^{P11}}}\right)$ if $\tilde{u}_{i0r}$=1

United States® Census Bureau

Modeling R&D Propensity

	Has R&D?	Second Collection (t =1)	
		Yes	No
First Collection (t = 0)	Yes		
$\tilde{u}_{i0r}$ = 0	No		

- Second Collection (t=1)
  - Linked sample data
  - Evidence of relationship between prior and current R&D status
- Conditional propensity
  - π_1=Prob(YES) at t = 1

Draw $\tilde{u}_{i1r} \sim Bernouilli\left(\frac{1}{1+e^{-\tilde{\alpha}_{0r}^{P01}}}\right)$ if $\tilde{u}_{i0r}$ = 0

United States® Census Bureau

Modeling R&D Expenditures ($$$)
Case 1: NO Prior R&D Expenditures

	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	Yes		
	No		

- Utilize weak linear relationship between characteristic and company size
- Annual Payroll Mixed Model
  - Random effects address year-to-year variation
  - ONLY fixed effects used for synthesis

United States® Census Bureau

Modeling R&D Expenditures
Case 1: NO Prior R&D Expenditures

	Has R&D?	**Synthesized Values of R&D Expenditures**
First Collection (t = 0)	Yes	
	No	

- Utilize weak linear relationship between characteristic and company size
- Annual Payroll Mixed Model
  - Random effects address year-to-year variation
  - ONLY fixed effects used for synthesis
- Use at
  - t = 0 (no historic data)

United States® Census Bureau

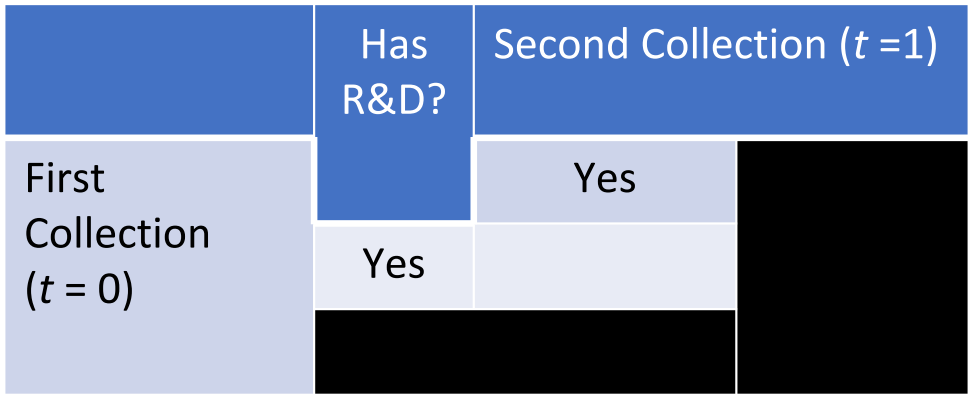
Modeling R&D Expenditures
Case 1: NO Prior R&D Expenditures

	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	No		

- Utilize weak linear relationship between characteristic and company size
- Annual Payroll Mixed Model
  - Random effects address year-to-year variation
  - ONLY fixed effects used for synthesis
- Use at
  - t = 0 (no historic data)
  - t =1 (no R&D expenditures at *t=0)*

United States® Census Bureau

Modeling R&D Expenditures
Case 2: Prior R&D Expenditures



- Utilize strong linear relationship between characteristic and previously reported value
- Prior R&D Expenditures Linear Model

United States® Census Bureau

Partially Synthetic Frames

- R = 1000 draws per company (= 1000 partially synthetic populations)
- Two time period values per company/draw
  - Frame variables from *t=0* and *t=1* (same in all synthetic populations)
  - Synthetic variables ($\tilde{u}_{i0r}$, $\tilde{u}_{i1r}$, $\tilde{y}_{i0r}$, $\tilde{y}_{i1r}$)

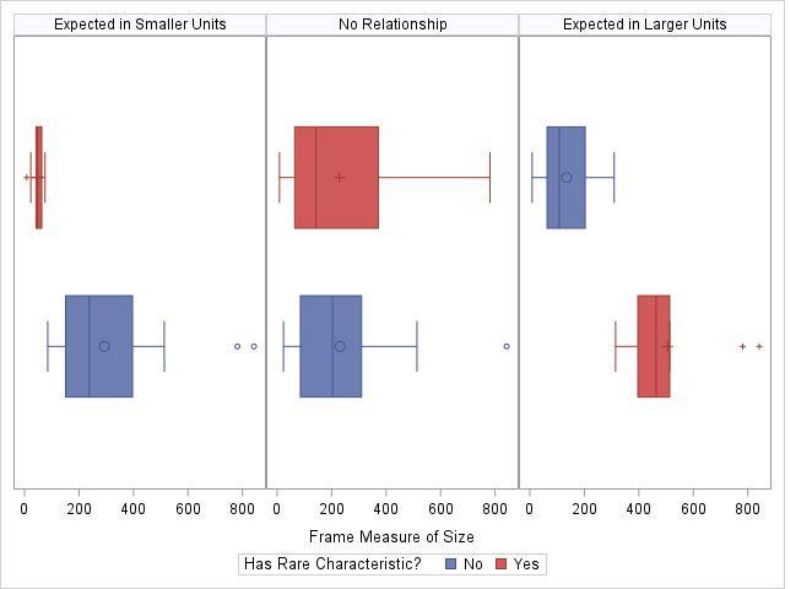
	Has R&D?	Second Collection (t =1)	
First Collection (t = 0)		Yes	No
	Yes		
	No		

United States® Census Bureau

Choose Synthetic Populations

- Generated 1,000 partially synthetic populations
  - Great for point estimation
  - Too many populations for repeated sample evaluation

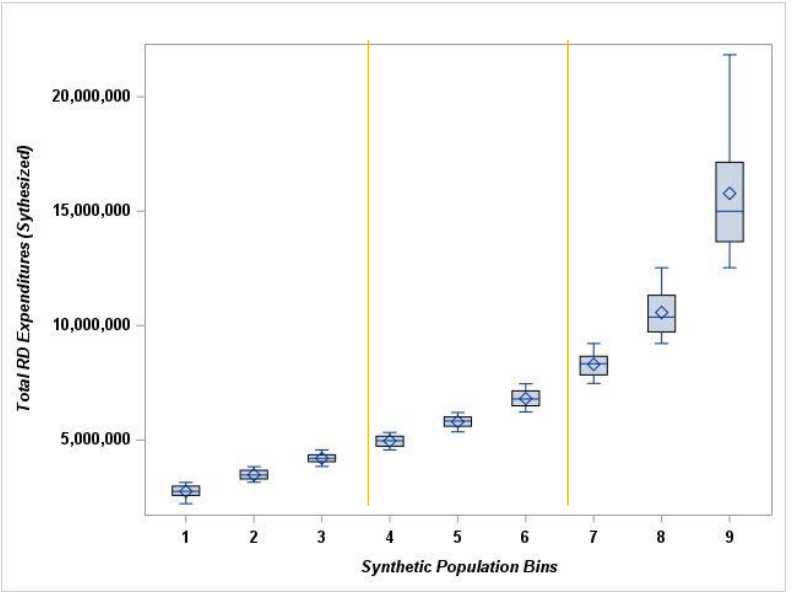
- Want robust sample design
  - Need a variety of partially synthetic populations!



United States® Census Bureau

Selection Procedure

- Group populations into 9 “bins” based on mean R&D expenditures (synthesized) at *t = 0*
  - Exclude outlying observations
  - *5th – 15th percentile,…,85th – 95th percentile*
- Randomly select one population from each bin



Smaller units more likely to have R&D

No apparent relationship

Larger units more likely to have R&D

At Last…Sampling (Case Study)

- Five industries
- Nine partially synthetic populations per industry
- Two scenarios
  - Case 1: No historic data (new survey)
  - Case 2: One prior (independent) sample
- Four candidate sample designs
- 500 independent samples per candidate sample design and synthetic population

United States® Census Bureau

At Last…Sampling (Case Study)

- Five industries
  - **Results from one industry presented**
- Nine partially synthetic populations per industry
- Two scenarios
  - Case 1: No historic data (new survey)
  - Case 2: One prior (independent) sample
- Four candidate sample designs
- 500 independent samples per candidate sample design and synthetic population

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

United States® Census Bureau

Candidate Sample Designs

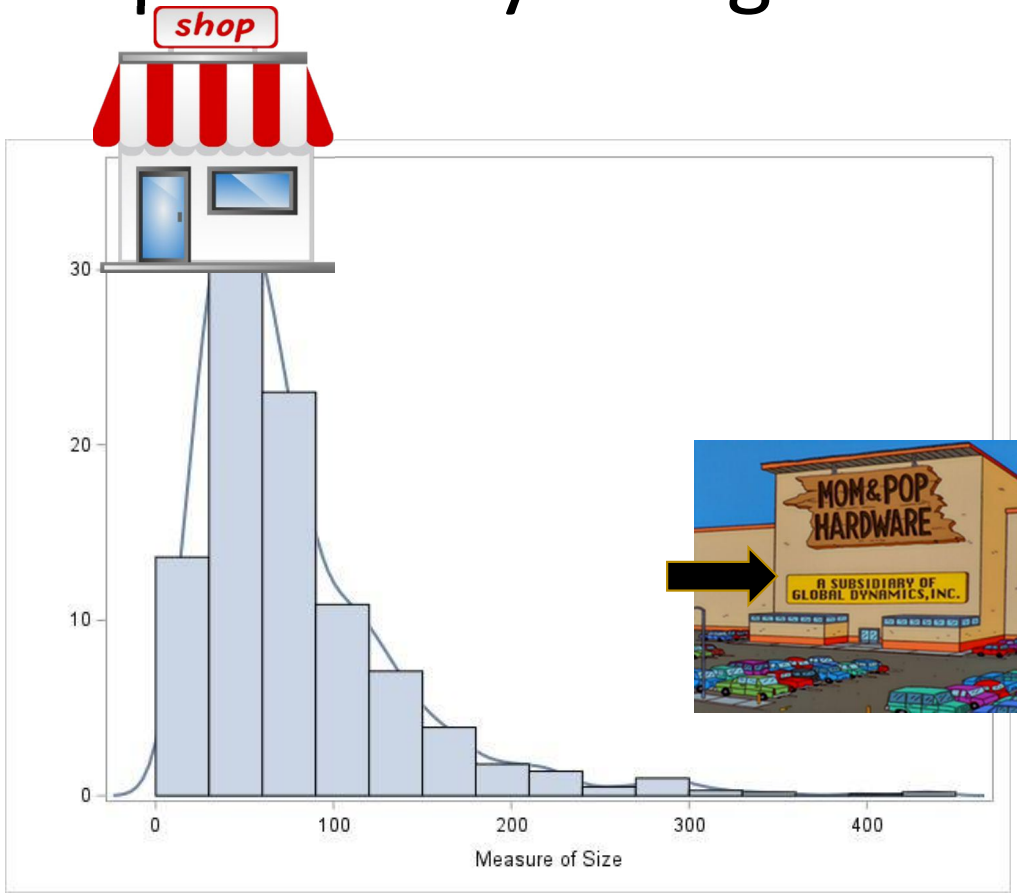
	Candidate Sample Design	Case 1: No Historic Data
SRS	Simple random sample without replacement	
PPS	Pareto sample without replacement	

Equal Versus Unequal Probability Sampling

United States® Census Bureau

Reminder: Business Sample Survey Designs

- Populations are skewed!
  - Small number of large companies
  - Majority small companies
- Publish TOTALS
  - And ratios of totals
- **Sample design must account for skewed population**



United States® Census Bureau