Towards Developing Synthetic Datasets for the Economic Census

Katherine Jenny Thompson*
Economic Statistical Methods Division
U.S. Census Bureau

Hang Kim
University of Cincinnati

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

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Economic Census Synthetic Data Project

- Current research is “proof of concept”
  - Synthetic industry-level micro-data
    - Subset of Economic Census Industries
    - Selected data items

- Ongoing project conducted by a team of methodologists and economists

- Builds on methodology developed and tested on one sector of the economy

Caveats

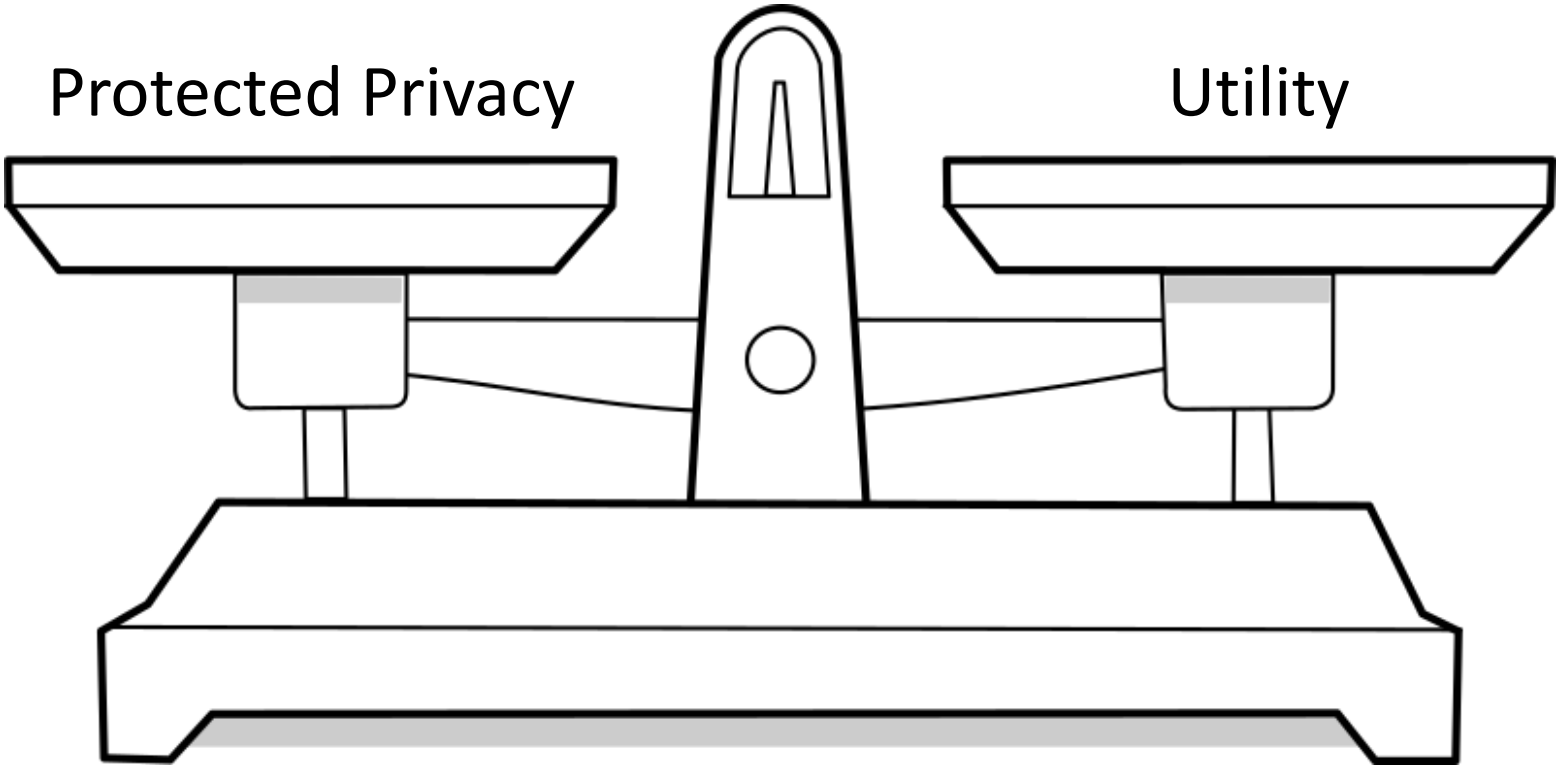
- Considerable progress in
  - Delineating the problem(s)
  - Defining our evaluation criteria
  - Obtaining necessary data and parameters, accessing computing environment, and running the code

- Anticipate presenting results in August 2018

Part 1: Introduction

Using simulated data in a fictional industry to prove my points

Overall Goal of Synthetic Data

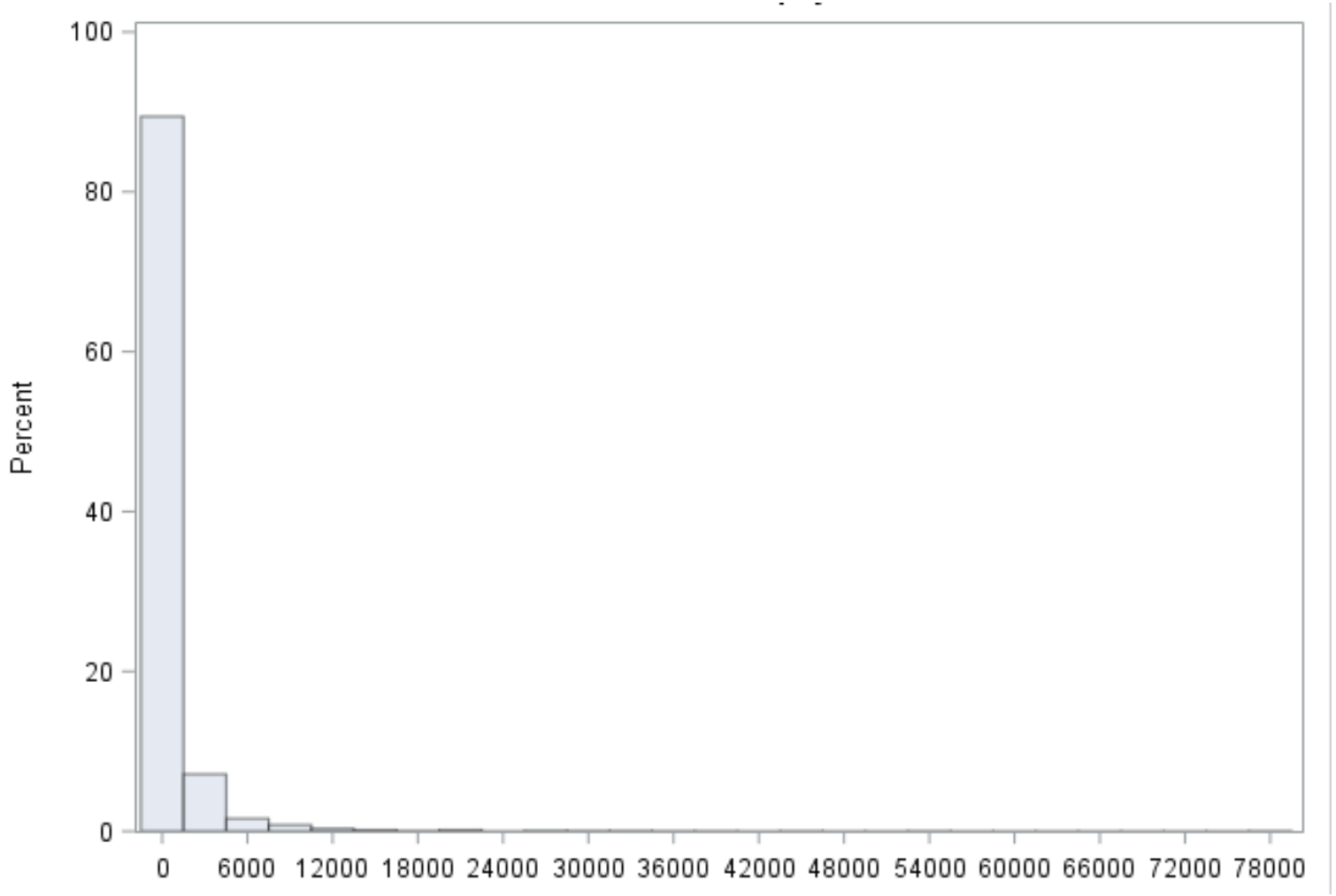


Defining Utility

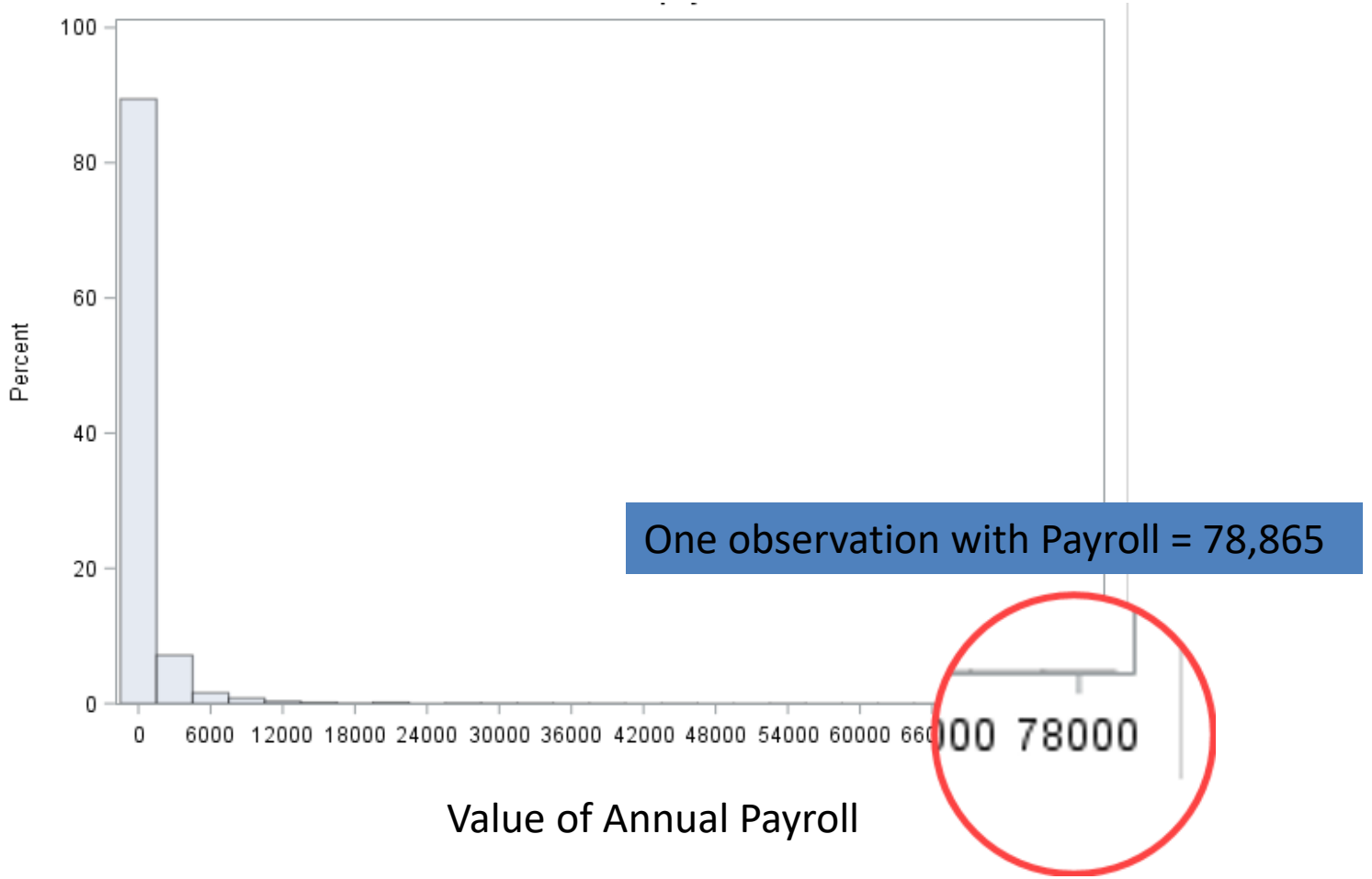
- Ultimate goal is develop multi-purpose synthetic micro-data to be shared with “public”
  - Economists, Statisticians, Data Users (Industry Experts)

- Conflicting desiderata
  - Economic models – multivariate relationships
  - Published benchmarks – aggregate totals

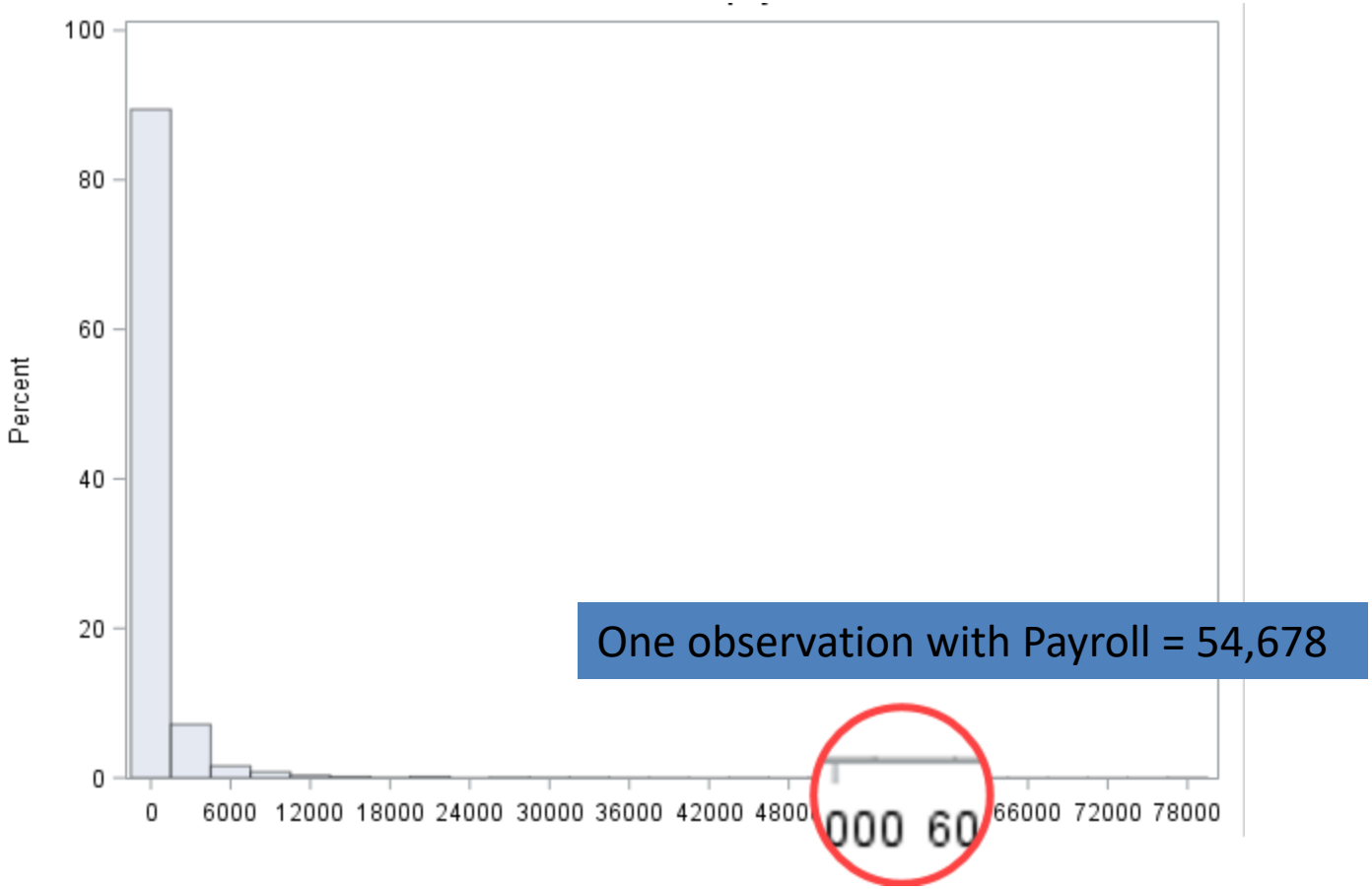
Privacy Versus Utility (Univariate)



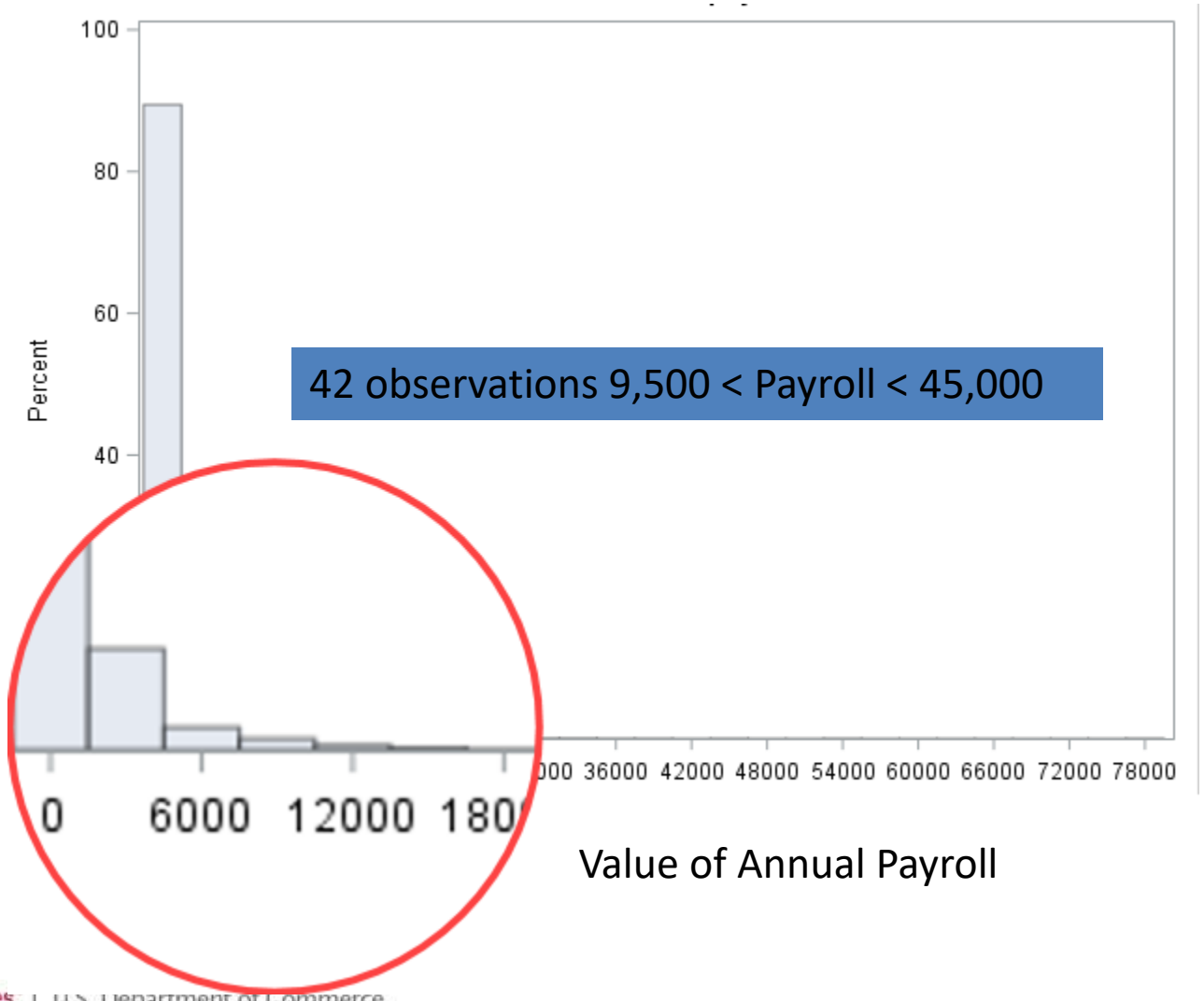
Privacy Versus Utility (Univariate)



Privacy Versus Utility (Univariate)



Privacy Versus Utility (Univariate)



Privacy Versus Utility (Univariate)

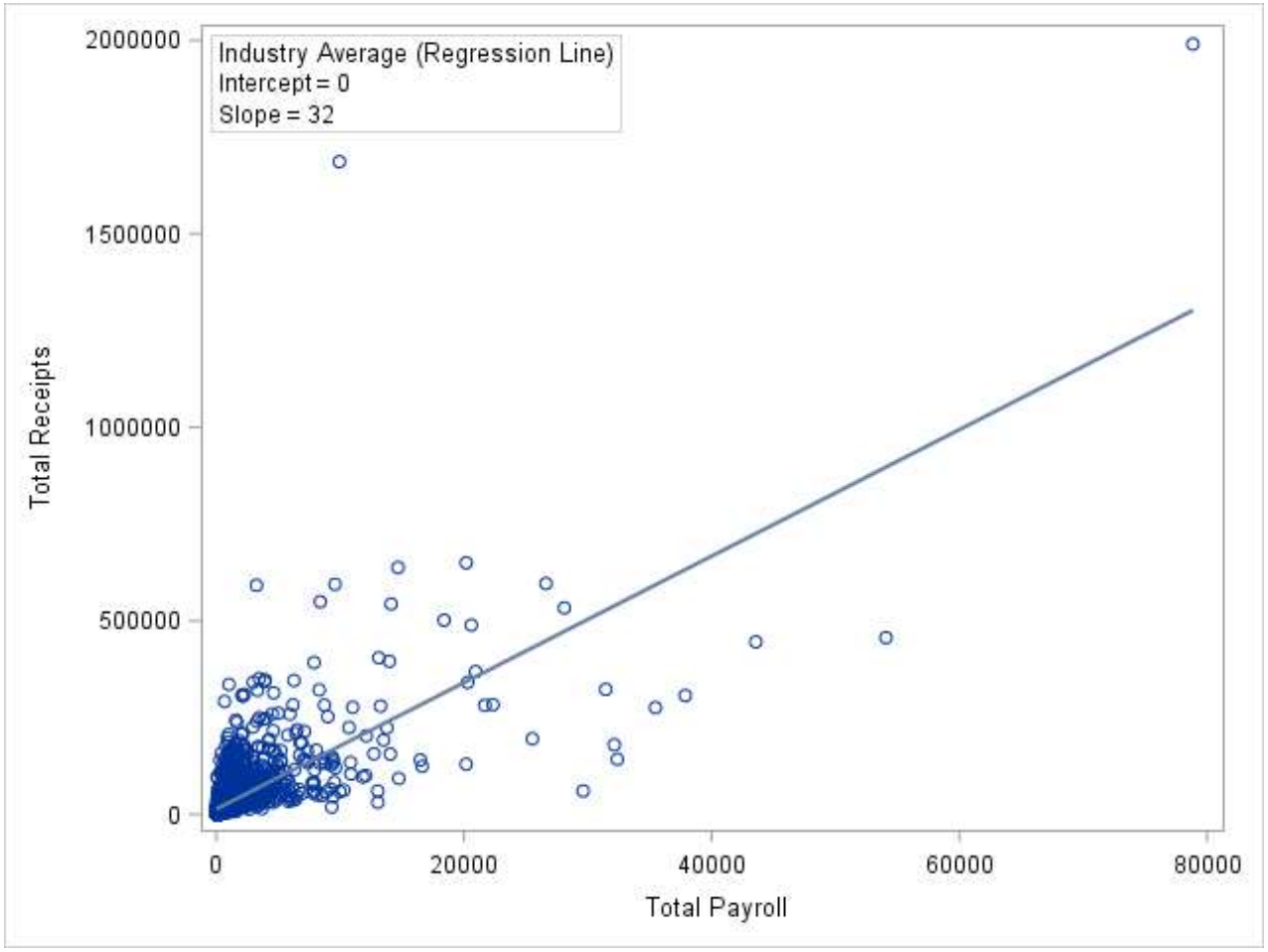
- Published “Industry”* Total
  - Total Payroll = 3,255,734
  - Total Payroll - Largest Observation = 3,176,869
  - Total Payroll - 2 Largest Observations = 3,122,791

- Largest businesses important for accurate totals (Utility)

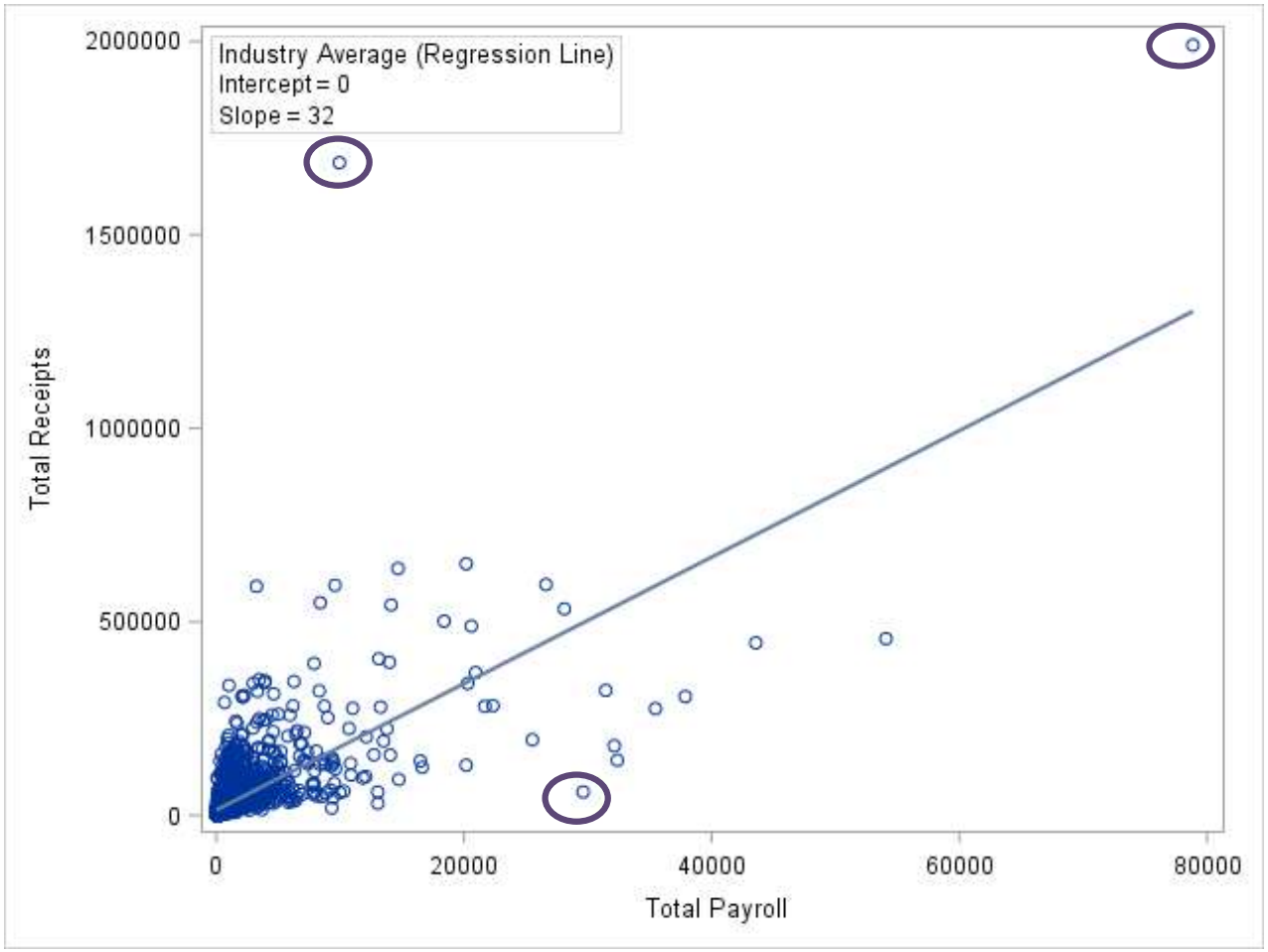
- But they need to be protected by mandate (Privacy)

* Fictional data ☺

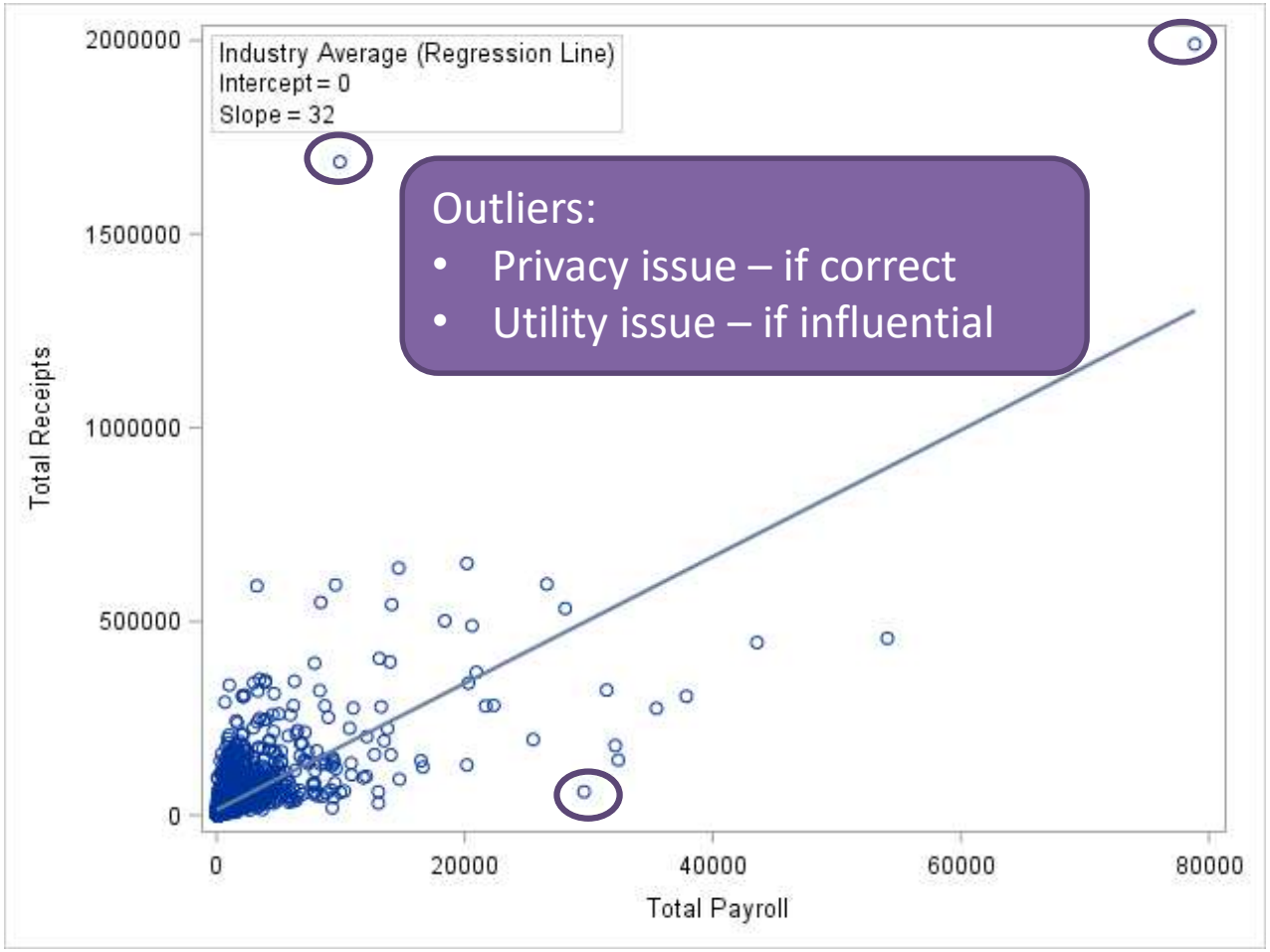
Privacy Versus Utility (Multivariate)



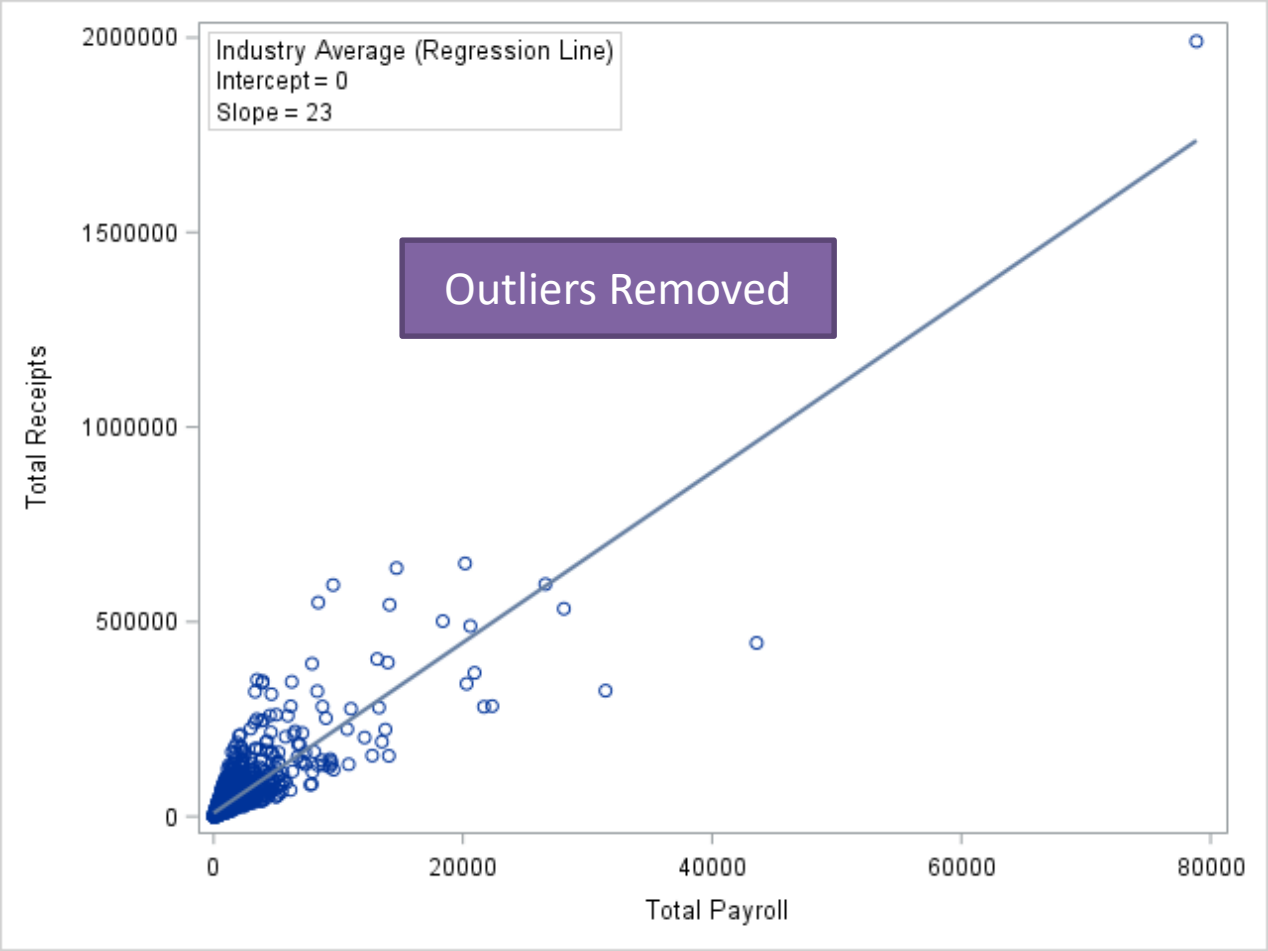
Privacy Versus Utility(Multivariate)



Privacy Versus Utility (Multivariate)



Privacy Versus Utility (Multivariate)



Privacy Versus Utility (Multivariate)

Privacy Versus Utility (Multivariate)

- Multivariate relationship obscured by outliers
  - Privacy concern – atypical profit or loss for industry
    - Potentially identifiable
  - Utility concern – influences estimation
    - Linear and nonlinear regression models

- But remember…conflicting desiderata
  - Univariate outliers are legitimate values

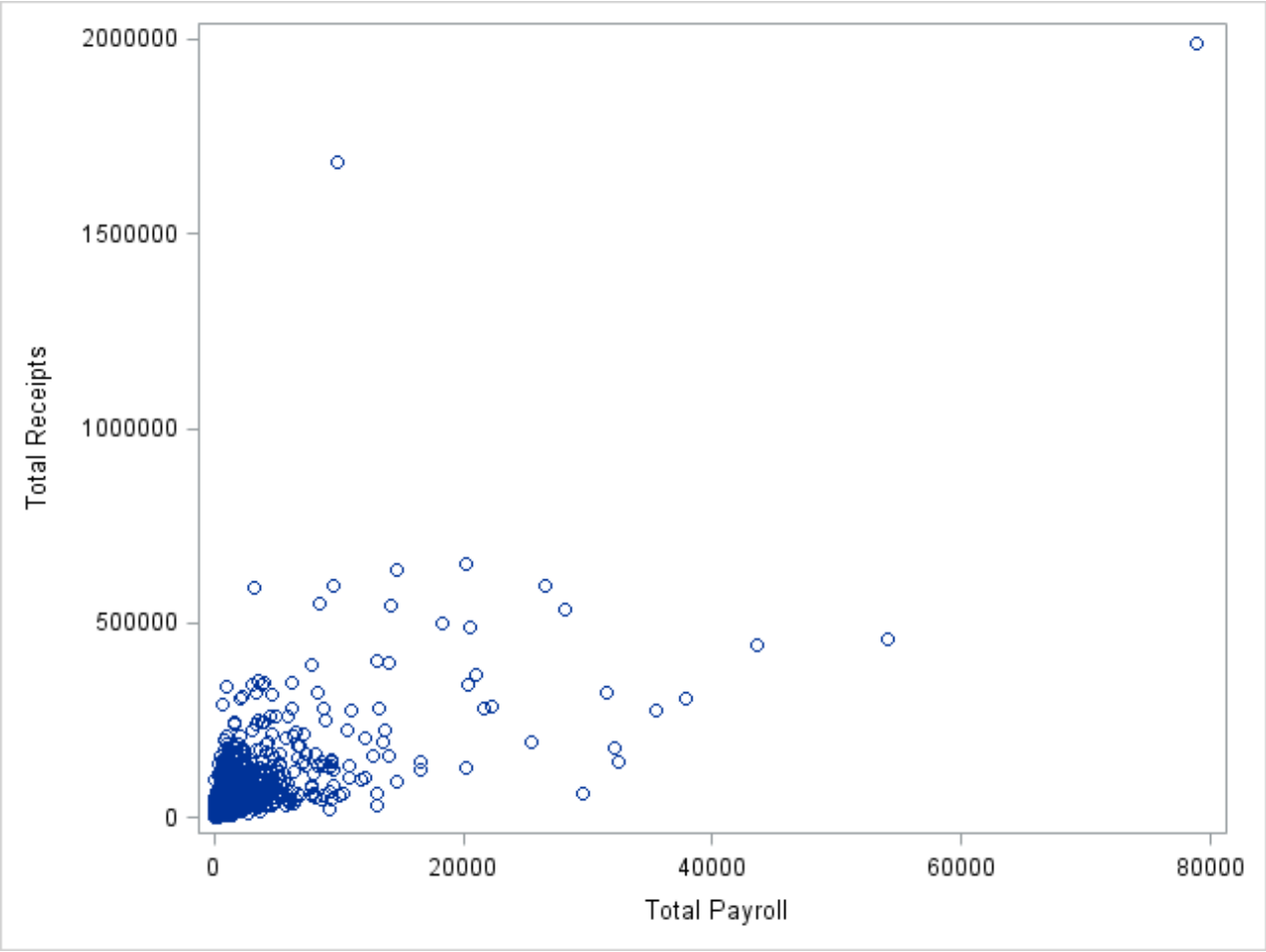
Economic Census

- Conducted every five years
- Collects data from establishments
- Covers all **economic** sectors except agriculture
  - North American **Industry** Classification System (NAICS)
  - 400+ industries
- Publishes totals (industry, industry by geography)

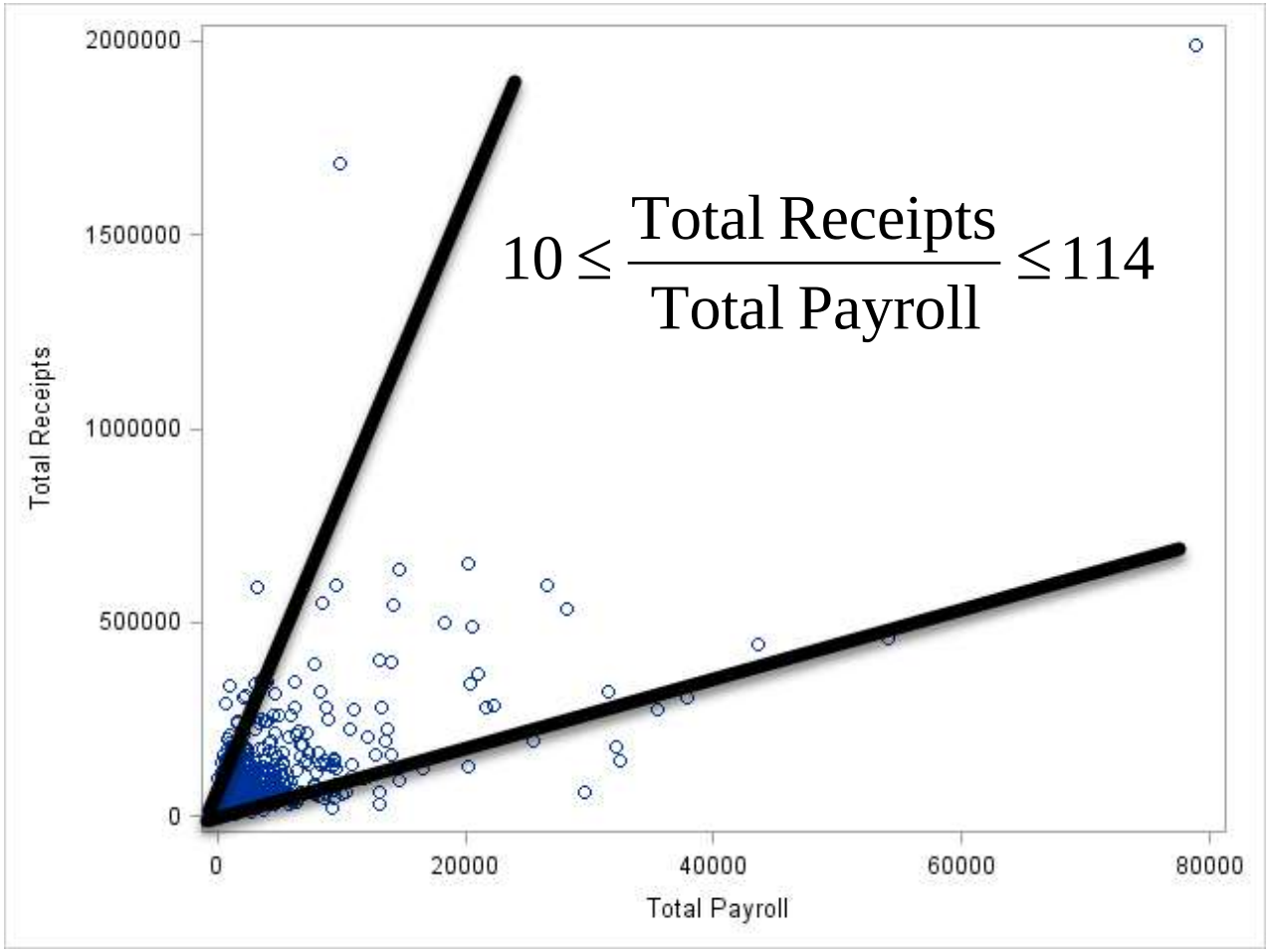
Economic Census

- Solicits
  - General Statistics (all units on frame)
  - Products and Special Inquiries (sampled units)
- Edit rules
  - Ratio edits
  - Range edits
  - Balance edits

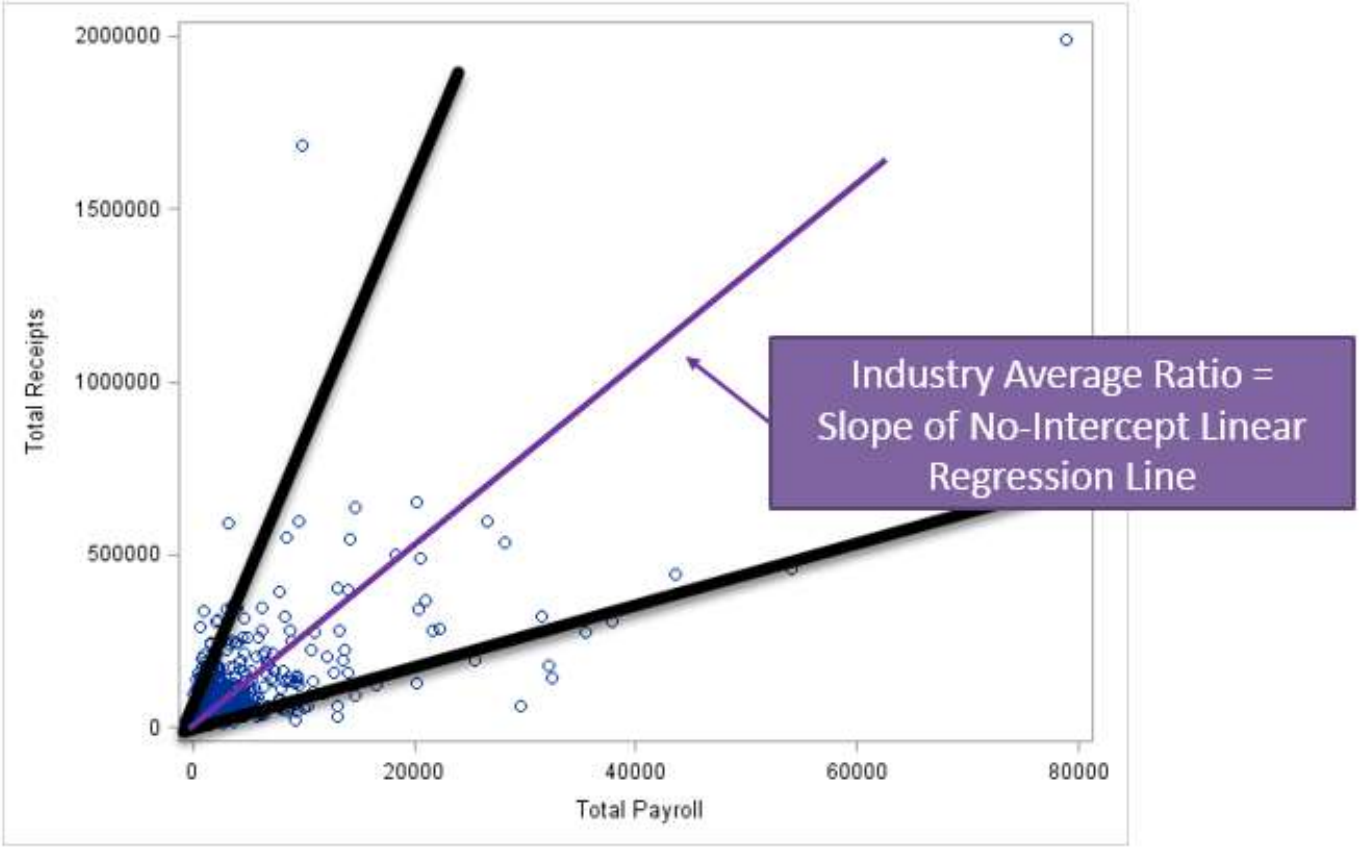
Ratio Edit (Multivariate)



Ratio Edit (Multivariate)



Ratio Edit (Multivariate)



- Ratio edits define no-intercept linear regression models
  - Numerator = Dependent variable
  - Denominator = Independent variable
- Generally use Weighted Least Squares for parameter estimation

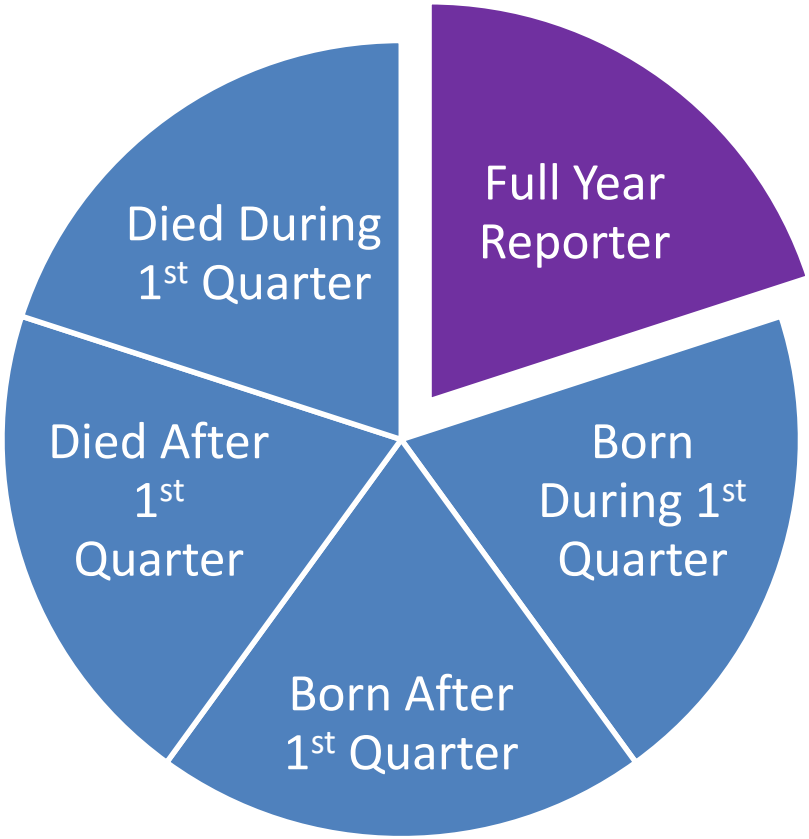
Economic Census: General Statistics Items

Sector	General Statistics Items	Total Items	Ratio Edits	Balance Edits
Finance, Insurance, Real Estate, Transportation, Communication, Utilities, Retail	Annual Payroll 1st Quarter Payroll 1st Quarter Employment Total Receipts	4	3	None
Services	Same + Operating Expenses (Tax-exempt establishments)	4 or 5	3 or 5	None
Wholesale Trade	Same + Cost of Purchases ,Operating Expenses , Beginning and Ending Inventories (Depending on Type of Establishment)	5 – 8	5-8	None
Manufacturing, Mining	Same + Cost of Materials (Details and Total), Benefits, Hours Worked, Beginning and Ending Inventories (Details and Total), Production Worker Wages and Other Employee wages	25	10	5
Construction	Not Discussed			

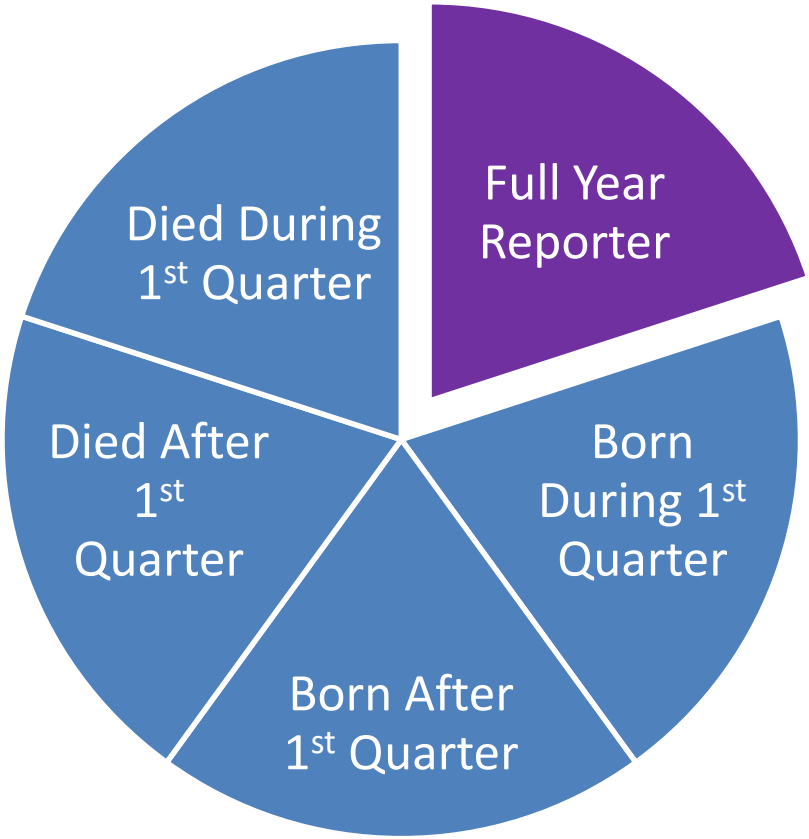
Economic Census: General Statistics Items

Sector	General Statistics Items	Total Items	Ratio Edits	Balance Edits
Finance, Insurance, Real Estate, Transportation, Communication, Utilities, Retail	Annual Payroll 1st Quarter Payroll 1st Quarter Employment Total Receipts	4	3	None
Services	Same + Operating Expenses (Tax-exempt establishments)	4 or 5	3 or 5	None
Wholesale Trade	Same + Cost of Purchases ,Operating Expenses , Beginning and Ending Inventories (Depending on Type of Establishment)	5 – 8	5-8	None
Manufacturing, Mining	Same + Cost of Materials (Details and Total), Benefits, Hours Worked, Beginning and Ending Inventories (Details and Total), Production Worker Wages and Other Employee wages	25	10	5
Construction	Not Discussed			

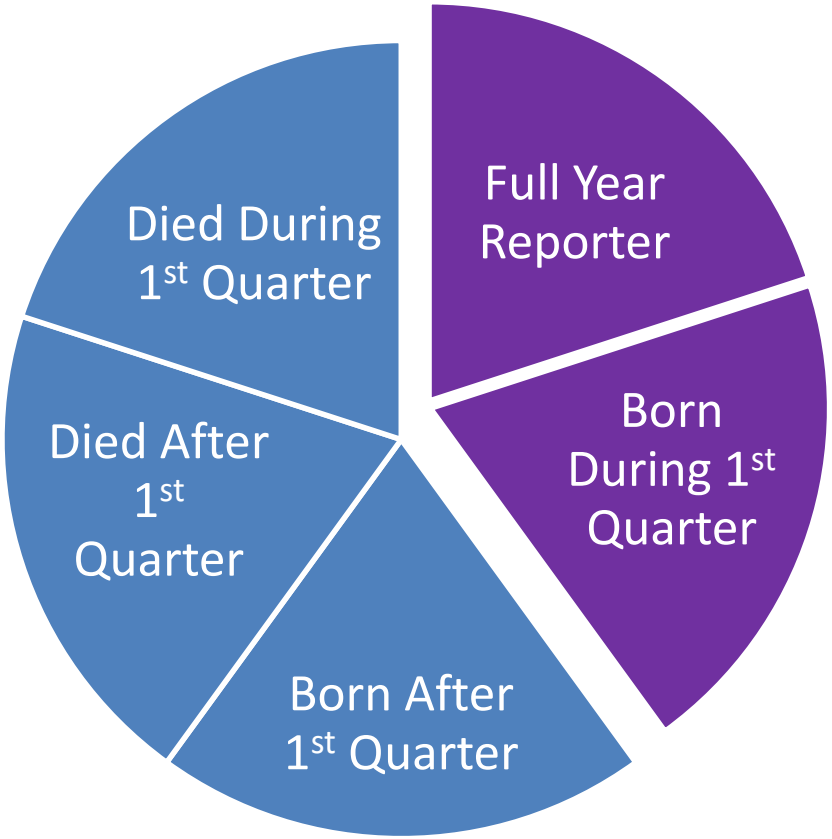
Economic Census: Types of Reporters



Economic Census: Types of Reporters

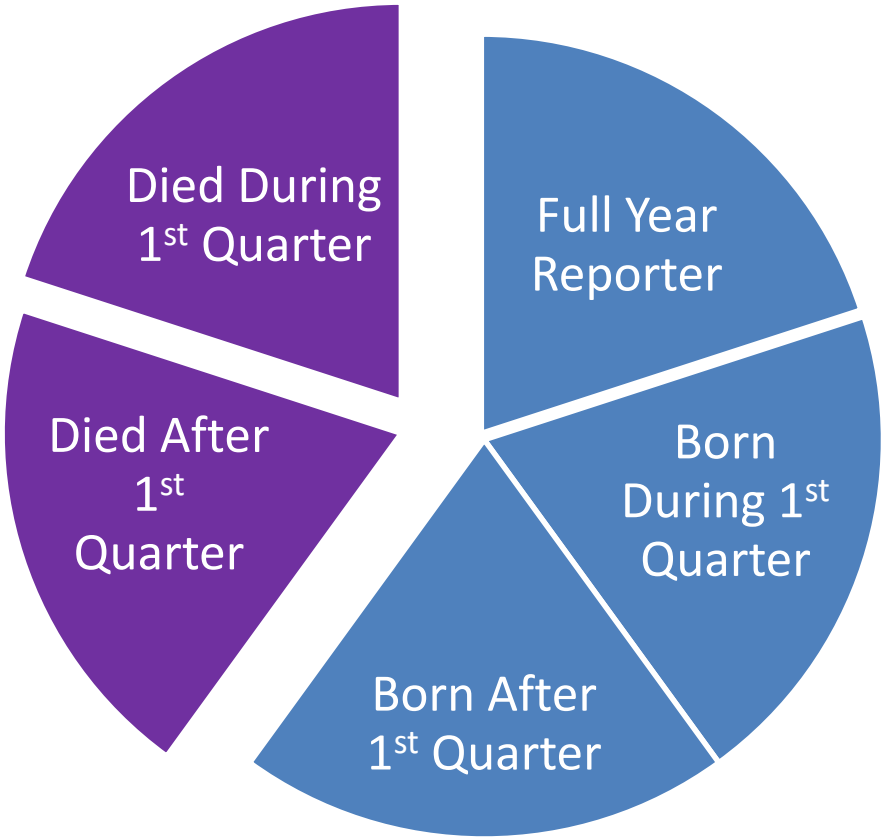


Economic Census: Types of Reporters



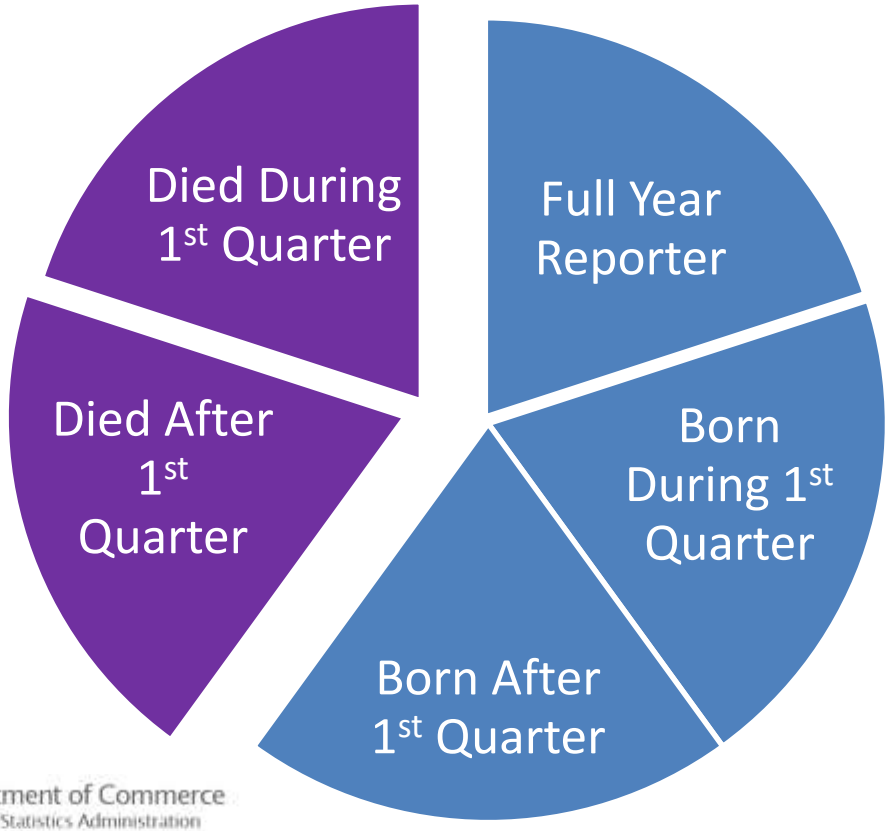
Consolidated for Modeling

Economic Census: Edit Modifications for Deaths



Economic Census: Edit Modifications for Deaths

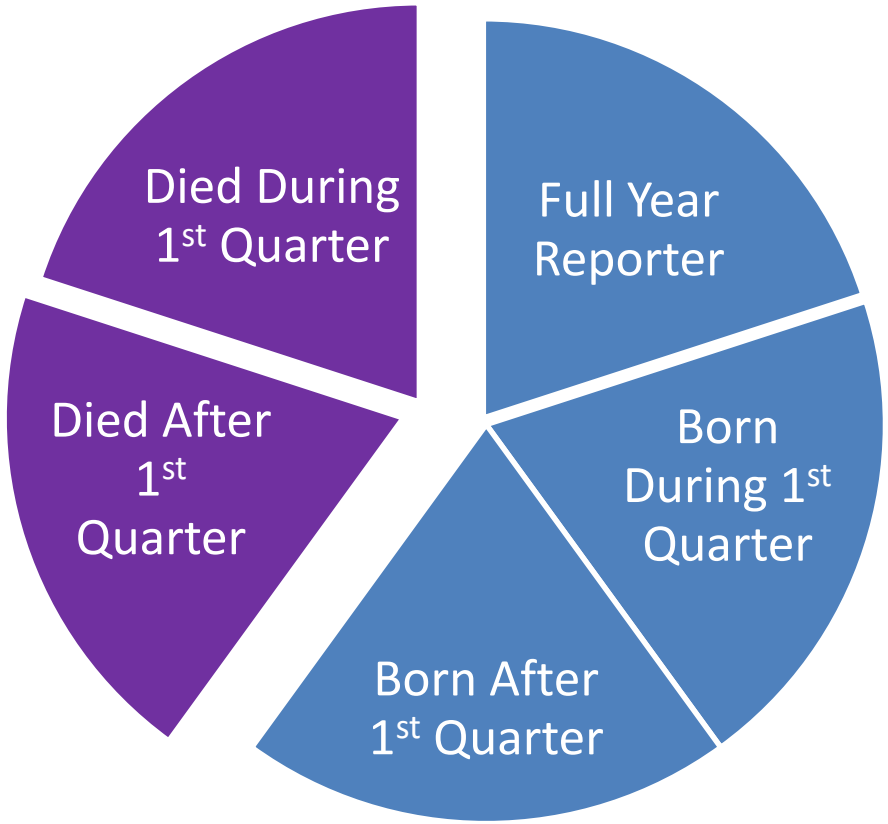
Prorate Upper Limits on Ratio Edits by Months in Business (Flow Items)



Economic Census: Edit Modifications for Deaths

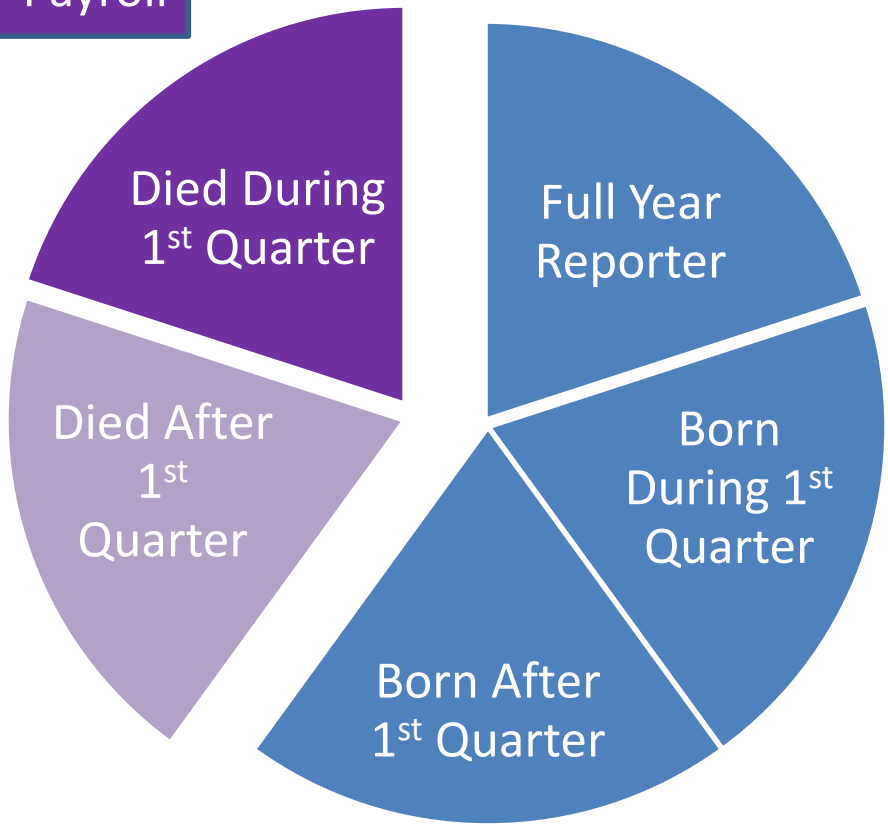
Prorate Upper Limits on Ratio Edits by Months in Business (Flow Items)

Reduce acceptable range for profits ratios



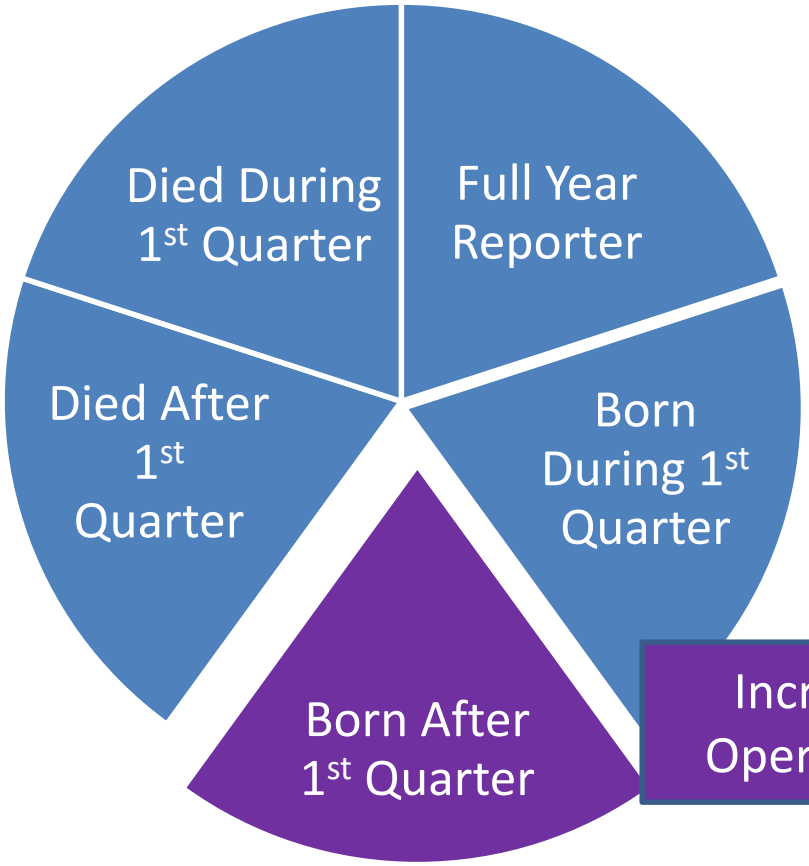
Economic Census: Edit Modifications for Deaths

Total Payroll = 1st Quarter Payroll



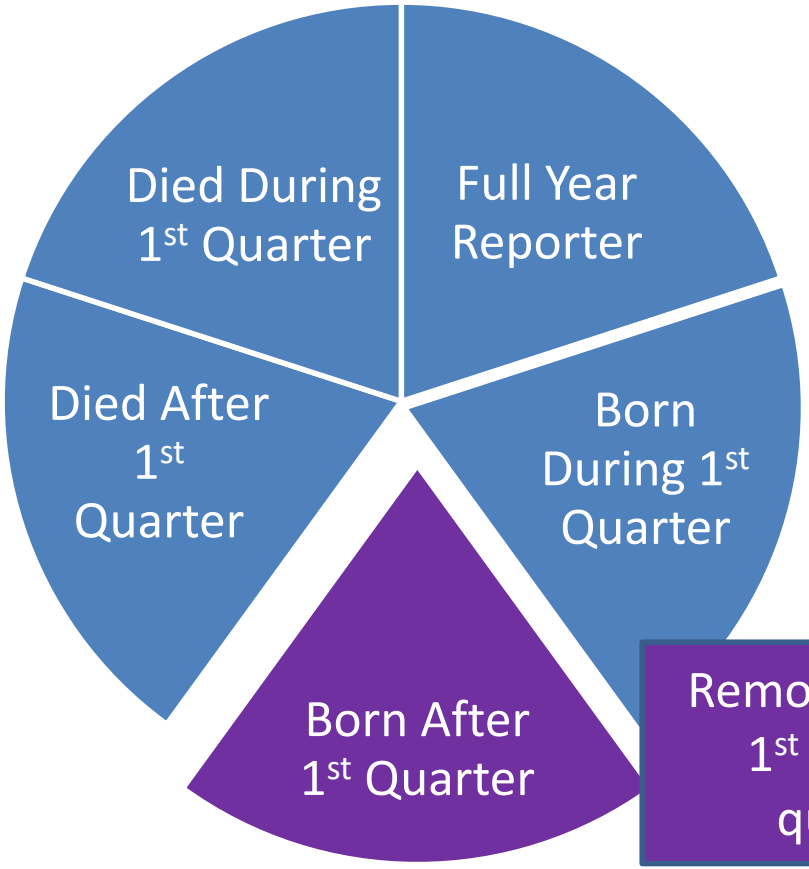
Economic Census: Edit Modifications for Births

Prorate Lower Limits on Ratio Edits by Months in Business (Flow Items)



Increase lower bound on Operating Expenses/Payroll

Economic Census: Edit Modifications for Births



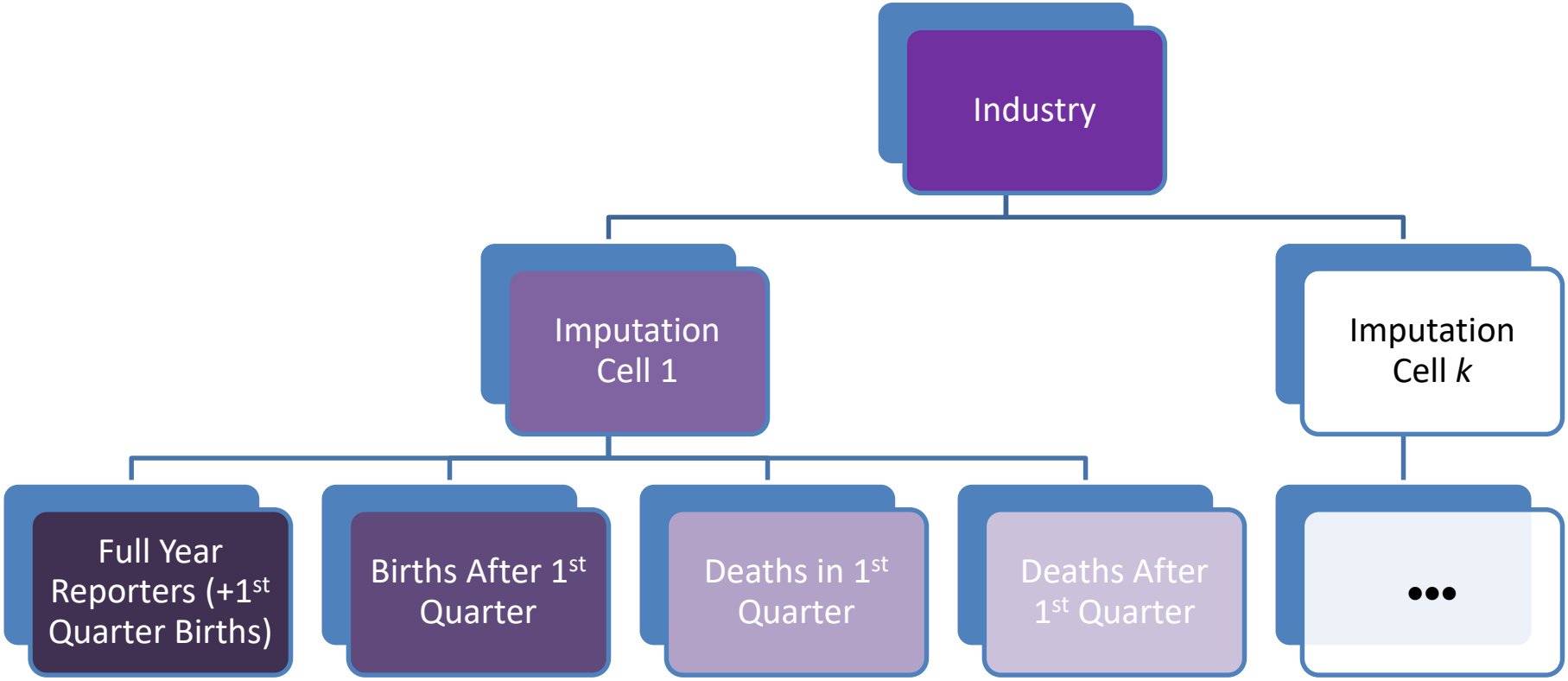
Remove all edits that contain 1st quarter payroll or 1st quarter employment

Part 2: Ongoing Research

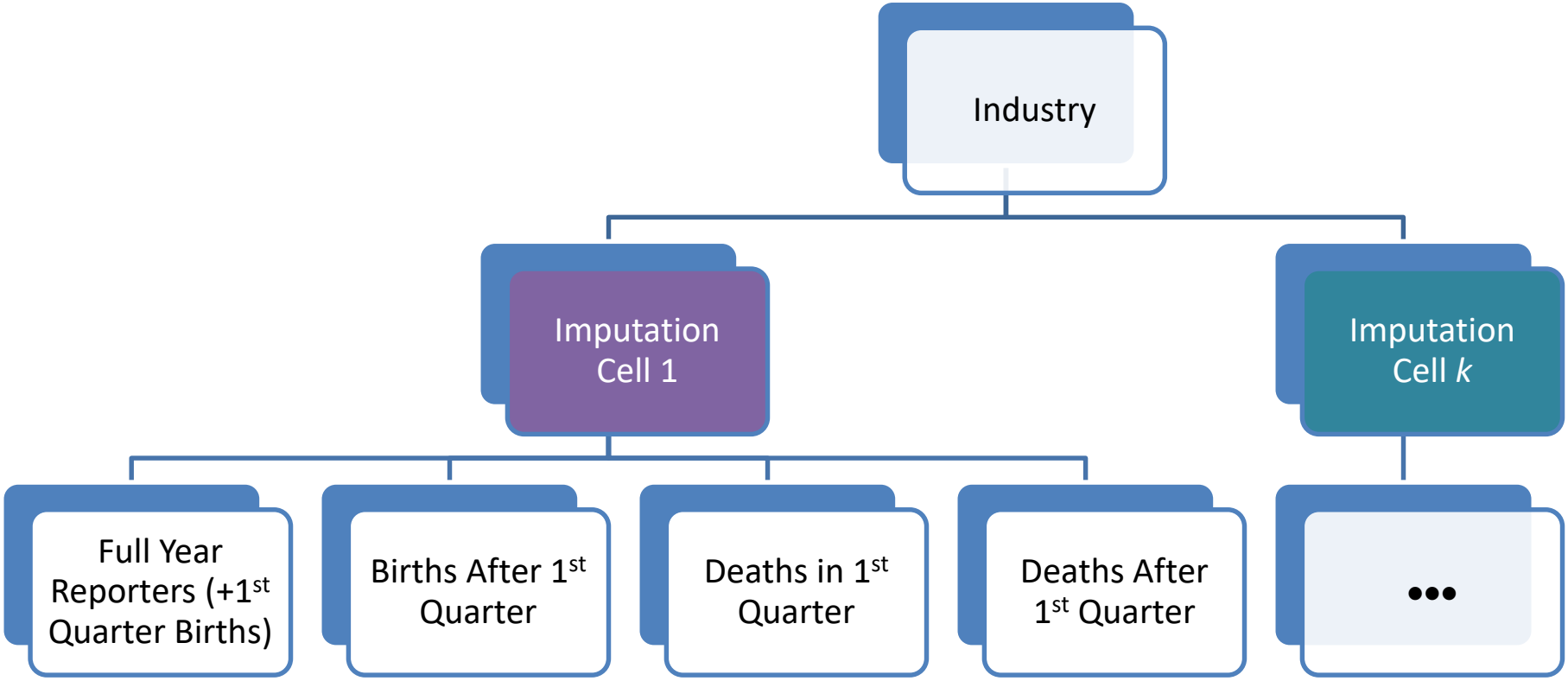
Economic Census Synthetic Data Project

- Synthetic industry-level micro-data (2012)
  - 53 Economic Census Industries provided
  - General statistics only
- Economic Census Edits
  - Ratio Edits
  - Subset of balance edits (software limitations)
- Full and part-year reporter distributions

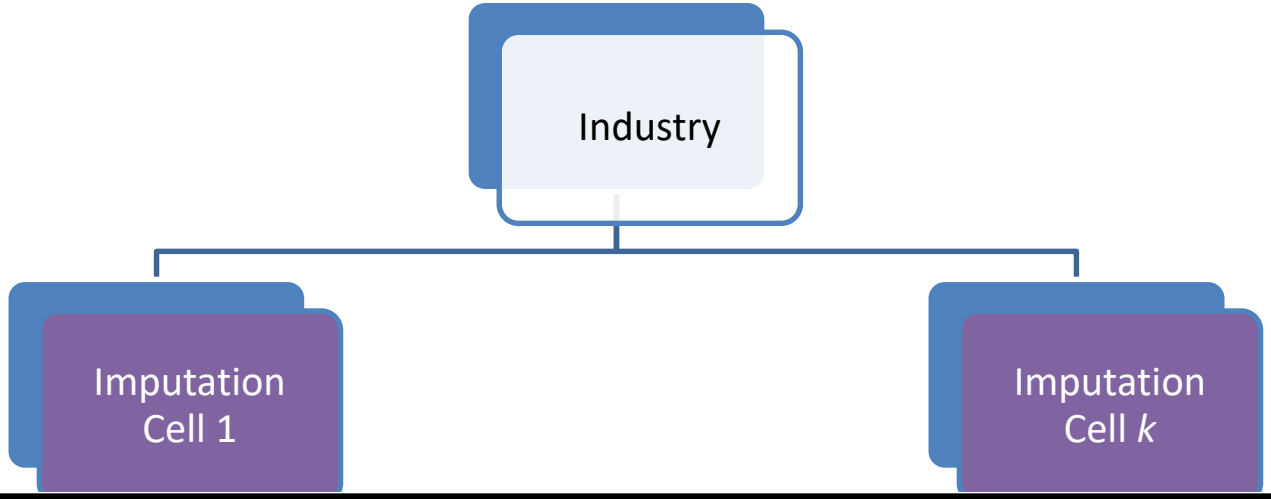
Procedure for Generating Economic Census Synthetic Data



Procedure for Generating Economic Census Synthetic Data

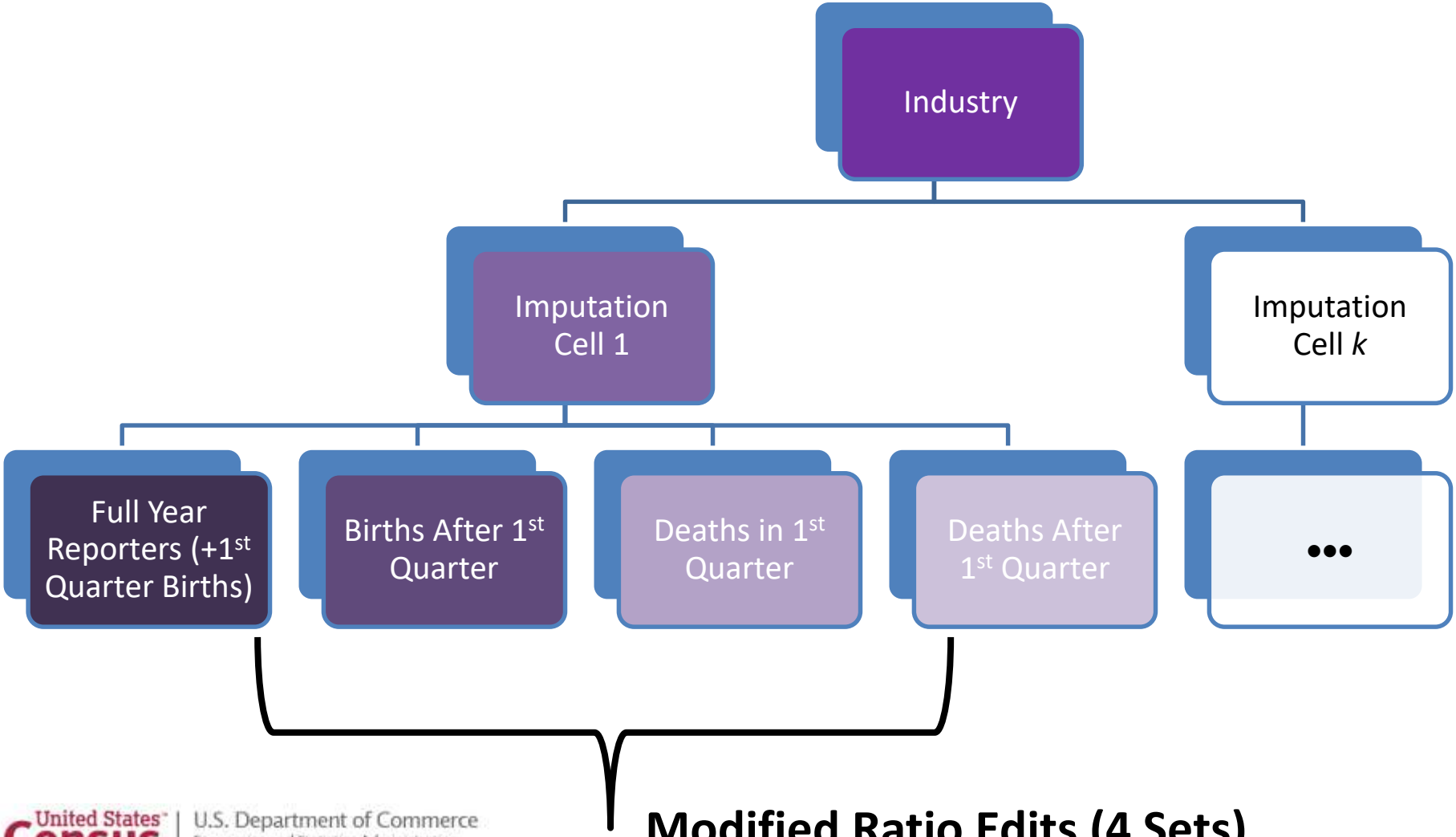


Procedure for Generating Economic Census Synthetic Data



Imputation Cell	Variables Collected							
Imputation Cell 1 **Brokers**	Annual Payroll	1st Quarter Payroll	1st Quarter Employment	Total Sales				
Imputation Cell 2 **Warehouses**	Annual Payroll	1st Quarter Payroll	1st Quarter Employment	Total Sales	Beginning Inventories	Ending Inventories	Cost of Purchases	Operating Expenses

Procedure for Generating Economic Census Synthetic Data



References (Proposed Methodology)

Kim, H. J., Reiter, J. P., and Karr, A. F. (2016). (online) Simultaneous edit-imputation and disclosure limitation for business establishment data. *Journal of Applied Statistics.*

Wang, Q., **Kim. H. J.** , Reiter, J. P., Cox, L. H., and Karr, A. F. (2016). EditImputeCont: Simultaneous Edit-Imputation for continuous microdata. R package version 1.0.1.

Kim, H. J., Karr, A. F., and Reiter, J. P. (2015). Statistical disclosure limitation in the presence of edit rules. *Journal of Official Statistics* 31, 121-138.

Kim, **H.J.,** Cox, L.F. , Karr, A.F. , Reiter, J.P., and Wang, Q. (2015). Simultaneous edit-imputation for continuous microdata, *Journal of the American Statistical Association* 110 , 987–999.

Methodology

Draw synthetic microdata $\boldsymbol{Y}^*_{Syn} = \{\boldsymbol{y}^*_1, \cdots, \boldsymbol{y}^*_n\}$ from its posterior predictive distribution:

$$f\left(\boldsymbol{Y}^*_{Syn} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right) = \int f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right) f\left(\boldsymbol{Y}^*_{Syn} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{T}, \boldsymbol{E}\right) d\boldsymbol{Y}_{EI}$$

where
$\widetilde{\boldsymbol{Y}}_O = \{\widetilde{\boldsymbol{y}}_1, \cdots, \widetilde{\boldsymbol{y}}_n\}$ denotes the original confidential microdata
$\boldsymbol{Y}_{EI} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ denotes the post edit-imputation microdata
$\boldsymbol{T} = (T_1, \cdots, T_p)$ the published totals in aggregate-level tables
$\boldsymbol{E}$ the edit rules.

Based on the model, the synthetic microdata $\boldsymbol{Y}^*_{Syn}$ are generated by two steps:

1. Draw $\boldsymbol{Y}_{EI}$ from $f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right)$ given the confidential and noisy microdata $\widetilde{\boldsymbol{Y}}_O$, then
2. Draw $\boldsymbol{Y}^*_{Syn}$ from $f\left(\boldsymbol{Y}^*_{Syn} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{T}, \boldsymbol{E}\right)$ given the post edit-imputation microdata $\boldsymbol{Y}_{EI}$.

Procedure for Generating Economic Census Data

Establishment	Annual Payroll	1st Quarter Payroll	1st Quarter Employment	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560		
4				
5	1026	257	15	15391
6	4819			
7				19472
…				
n	1430	357	20	21449

Procedure for Generating Economic Census Data

- One Imputation Cell
- One Type of Reporter
- Missing and Erroneous Data

		1st Quarter Payroll	1st Quarter Employment	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560		
4				
5	1026	257	15	15391
6	4819			
7				19472
…				
n	1430	357	20	21449

Methodology: Step 1

Draw synthetic microdata $\boldsymbol{Y}^*_{Syn} = \{\boldsymbol{y}^*_1, \cdots, \boldsymbol{y}^*_n\}$ from its posterior predictive distribution:

$$f\left(\boldsymbol{Y}^*_{Syn} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right) = \int f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right) f\left(\boldsymbol{Y}^*_{Syn} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{T}, \boldsymbol{E}\right) d\boldsymbol{Y}_{EI}$$

where
$\widetilde{\boldsymbol{Y}}_O = \{\widetilde{\boldsymbol{y}}_1, \cdots, \widetilde{\boldsymbol{y}}_n\}$ denotes the original confidential microdata
$\boldsymbol{Y}_{EI} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ denotes the post edit-imputation microdata
$\boldsymbol{T} = (T_1, \cdots, T_p)$ the published totals in aggregate-level tables
$\boldsymbol{E}$ the edit rules.

Based on the model, the synthetic microdata $\boldsymbol{Y}^*_{Syn}$ are generated by two steps:

1. Draw $\boldsymbol{Y}_{EI}$ from $f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right)$ given the confidential and noisy microdata $\widetilde{\boldsymbol{Y}}_O$, then
2. Draw $\boldsymbol{Y}^*_{Syn}$ from $f\left(\boldsymbol{Y}^*_{Syn} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{T}, \boldsymbol{E}\right)$ given the post edit-imputation microdata $\boldsymbol{Y}_{EI}$.

Procedure for Generating Economic Census Data

Original Data

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560		
4				
5	1026	257	15	15391
6	4819			
7				19472
…				
n	1430	357	20	21449

Multiply Imputed Data (Implicate 1)

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560	89	93579
4	2160	123	129606	2160
5	1026	257	15	15391
6	4819	1205	69	72291
7	1298	325	19	19472
…				
n	1430	357	20	21449

Procedure for Generating Economic Census Data

Original Data

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560		
4				
5	1026	257	15	15391
6	4819			
7				19472
…				
n	1430	357	20	21449

Multiply Imputed Data (Implicate 1)

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560	89	93579
4	2160	123	129606	2160
5	1026	257	15	15391
6	4819	1205		
7	1298	325		
…				
n	1430	357	20	21449

Four more multiply imputed datasets (implicates)

Procedure for Generating Economic Census Data

Original Data

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560		
4				
5	1026	257	15	15391
6	4819			
7				19472
…				
n	1430	357	20	21449

Multiply Imputed Data (Implicate 1)

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560	89	93579
4	2160	123	129606	2160
5	1026	257	15	15391
6	4819	1205		
7	1298	325		
…				
n				

Four more multiply imputed datasets (implicates)

Procedure repeated FOUR times per imputation cell (by type of reporter)

Methodology

Draw synthetic microdata $\boldsymbol{Y}_{Syn}^{*} = \{\boldsymbol{y}_1^{*}, \cdots, \boldsymbol{y}_n^{*}\}$ from its posterior predictive distribution:

$$f\left(\boldsymbol{Y}_{Syn}^{*} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right) = \int f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right) f\left(\boldsymbol{Y}_{Syn}^{*} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{T}, \boldsymbol{E}\right) d\boldsymbol{Y}_{EI}$$

where
$\widetilde{\boldsymbol{Y}}_O = \{\widetilde{\boldsymbol{y}}_1, \cdots, \widetilde{\boldsymbol{y}}_n\}$ denotes the original confidential microdata
$\boldsymbol{Y}_{EI} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n\}$ denotes the post edit-imputation microdata
$\boldsymbol{T} = (T_1, \cdots, T_p)$ the published totals in aggregate-level tables
$\boldsymbol{E}$ the edit rules.

Based on the model, the synthetic microdata $\boldsymbol{Y}_{Syn}^{*}$ are generated by two steps:

1. Draw $\boldsymbol{Y}_{EI}$ from $f\left(\boldsymbol{Y}_{EI} \middle| \widetilde{\boldsymbol{Y}}_O, \boldsymbol{T}, \boldsymbol{E}\right)$ given the confidential and noisy microdata $\widetilde{\boldsymbol{Y}}_O$, then
2. Draw $\boldsymbol{Y}_{Syn}^{*}$ from $f\left(\boldsymbol{Y}_{Syn}^{*} \middle| \boldsymbol{Y}_{EI}, \boldsymbol{T}, \boldsymbol{E}\right)$ given the post edit-imputation microdata $\boldsymbol{Y}_{EI}$.

Procedure for Generating Economic Census Data

Multiply Imputed Data (Implicate 1)

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560	89	93579
4	2160	123	129606	2160
5	1026	257	15	15391
6	4819	1205	69	72291
7	1298	325	19	19472
…				
n	1430	357	20	21449

Synthetic Data
1st Implicate Produced from Implicate 1

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	9250	2312	132	138749
2	2406	601	34	36089
3	1560	390	22	23395
4	3888	972	56	58323
5	154	38	2	2309
6	578	145	8	8675
7	428	107	6	6426
…				
n	543	136	8	8151

Procedure for Generating Economic Census Data

Multiply Imputed Data (Implicate 1)

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560	89	93579
4	2160	123	129606	2160
5	1026	257	15	15391
6	4819	1205	69	72291
7	1298	325	19	19472
…				
n	1430	357	20	21449

Synthetic Data
1st Implicate Produced from Implicate 1

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	9250	2312	132	138749
2	2406	601	34	3608
3	1560	390	22	23395
4	3888	972	56	58323
5	154			
6	578			
7	428	107	6	6426
…				
n	543	136	8	8151

Four more multiply imputed datasets (implicates) drawn from Implicate 1

Procedure for Generating Economic Census Data

Multiply Imputed Data (Implicate 1)

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	7283	1821	104	109251
2	6874	1719	98	103112
3	6239	1560	89	93579
4	2160	123	129606	2160
5	1026	257	15	15391
6	4819	1205	69	72291
7	1298	325	19	19472
…				
n	1430	357	20	21449

Synthetic Data
1st Implicate Produced from Implicate 1

Unit	Annual Payroll	1st Quarter Payroll	1st Quarter Emp.	Total Sales
1	9250	2312	132	138749
2	2406	601	34	3608
3	1560	390	22	23395
4	3888	972	56	58323
5	154			
6	578			
7	428	107	6	6426

Four more multiply imputed datasets (implicates) drawn from Implicate 1

Procedure repeated FOUR times per Implicate/Imputation Cell (by type of reporter)

Properties of Synthetic Data

- Preserves multivariate relationships
  - Does not preserve marginal moments of reported data that satisfies all edit constraints
  - Proposed model enhancement to preserve totals

- Privacy protected by preserving the joint distributions between items but altering (not preserving) original reported values of units

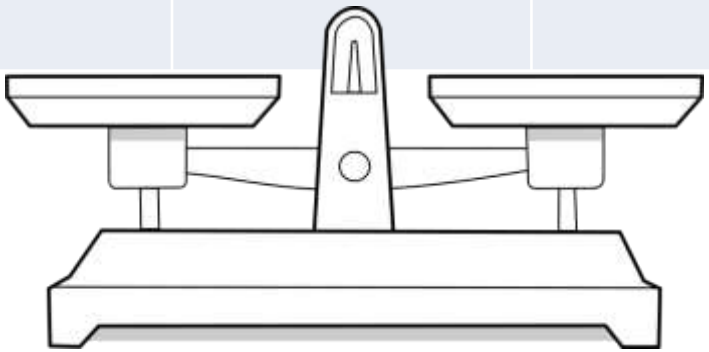
Protected Privacy Vs. Utility

	Protected Privacy	Utility
Editing/Imputation Procedure All items equally likely to be erroneous	–	↓
Different Models by Type Of Unit Full Year/Part Year	↓	↑
Multiple Imputation (Editing/Imputation)	↑	↑
Multiple Draws from Posterior Predictive Distribution (Synthetic Data)	↑	↓

Protected Privacy Vs. Utility

	Protected Privacy	Utility
Editing/Imputation Procedure All items equally likely to be erroneous	–	↓
Different Models by Type Of Unit Full Year/Part Year	↓	↑
Multiple Imputation (Editing/Imputation)	↑	↑
Multiple Draws from Posterior Predictive Distribution (Synthetic Data)	?	↓

Protected Privacy



Utility

Quality (Utility) Metrics

- Multivariate relationships (Primary)
  - Correlations between all items
  - Weighted Least Squares Regression Parameters
    - Ratio edit pairs ("Industry Averages")
  - Logistic regression model fit and regression parameters
    - Dependent variable = death (of business)
- Univariate moments (Secondary)

NOTE: *"Gold Standard" values obtained from reported data that satisfies all ratio edits*

Confidentiality/Disclosure Avoidance Metrics

- Formal protection of largest company value for multiply-imputed totals (all variables)
  - Estimate derived by subtraction
  - Bounded by fixed percentage (TBD)

- (Possible) Formal protection of largest synthesized establishment values in synthetic data implicates

Current Approach (Testing)

In each test industry

- Run Editing/Imputation programs separately
  - Imputation Cells/Type of Reporter
- Resample to create synthetic data
  - Imputation Cell/Type of reporter
- Consolidate data to industry level
- Utility and Confidentiality Assessment
  - Industry
  - Imputation Cell by Type of Reporter

Proposed Enhancements (Existing Models)

- Refine edit-imputation models
  - Hierarchical model for full/part year reporters
  - Incorporate informative priors that reflect item reliability
  - Identify/correct rounding errors
- Synthetic data
  - Calibration constraint addressed by adding penalty function for difference between the aggregated synthetic data items and the marginal total

Final Comments

- Deeper understanding of the nuances of the input data and the synthetic data requirements
  - Temporary workarounds
  - Necessary model refinements
- Modeling census data
  - Need to address sampling issues

Synthetic Data Research Team

Jenny Thompson
Hang Kim*
Aref Dajani
Charles Coleman
Daniel Whitehead
Eric Valentine
Kevin Bembridge
Kirk White

Maria Garcia
Nelson Chung
Noah Bassel
Phyllis Singer
Steven Riesz
Yarissa Gonzalez

*University of Cincinnati