Data Synthesis and Perturbation for the American Community Survey at the U.S. Census Bureau

Amy Lauger, U.S. Census Bureau

Michael Freiman, U.S. Census Bureau

Jerry Reiter, Duke University and U. S. Census Bureau

2016 Joint Statistical Meetings

Dual Initiatives at the Census Bureau

- Data Dissemination Transformation
  - Easier to access
  - More flexible and customizable
  - Combines multiple data sources into integrated data products
- Disclosure Avoidance Transformation
  - Increase in publicly available data and sophistication of data mining techniques
  - To facilitate transformation of data dissemination

New Methods Research

- Two ways to protect confidentiality
  - Suppress data
  - Perturb data
- Microdata Analysis System (Discontinued)
  - Remote table server with official microdata as source
  - **Table suppression** as primary DA method
  - Discontinued research due to low data utility and unacceptable disclosure risk
- Expanded use of synthetic data
- Formal privacy mechanisms and criteria

Current Synthetic Products

- Decennial Census and American Community Survey (ACS): group quarters data
- Survey of Income and Program Participation Synthetic Beta
- Synthetic Longitudinal Business Database
- OnTheMap

ACS Household Data Synthesis: Overarching Questions

- Can we provide high quality synthetic data that adequately preserve relationships across all ACS variables?
- Which models should we use?
- At which geographic level should we synthesize?
- Should we do partial or full synthesis?
- How best to incorporate survey weights?

Research Phase 1

- Create synthetic datasets for:
  - Sex, Age, Race, Hispanic Origin, Educational Attainment, Marital Status, and Wages
- Use ACS public-use microdata (PUMS)
  - Simulate wages to undo rounding and topcoding
- Use a relatively small geography (n=2500)
- Use Classification and Regression Trees (CART)

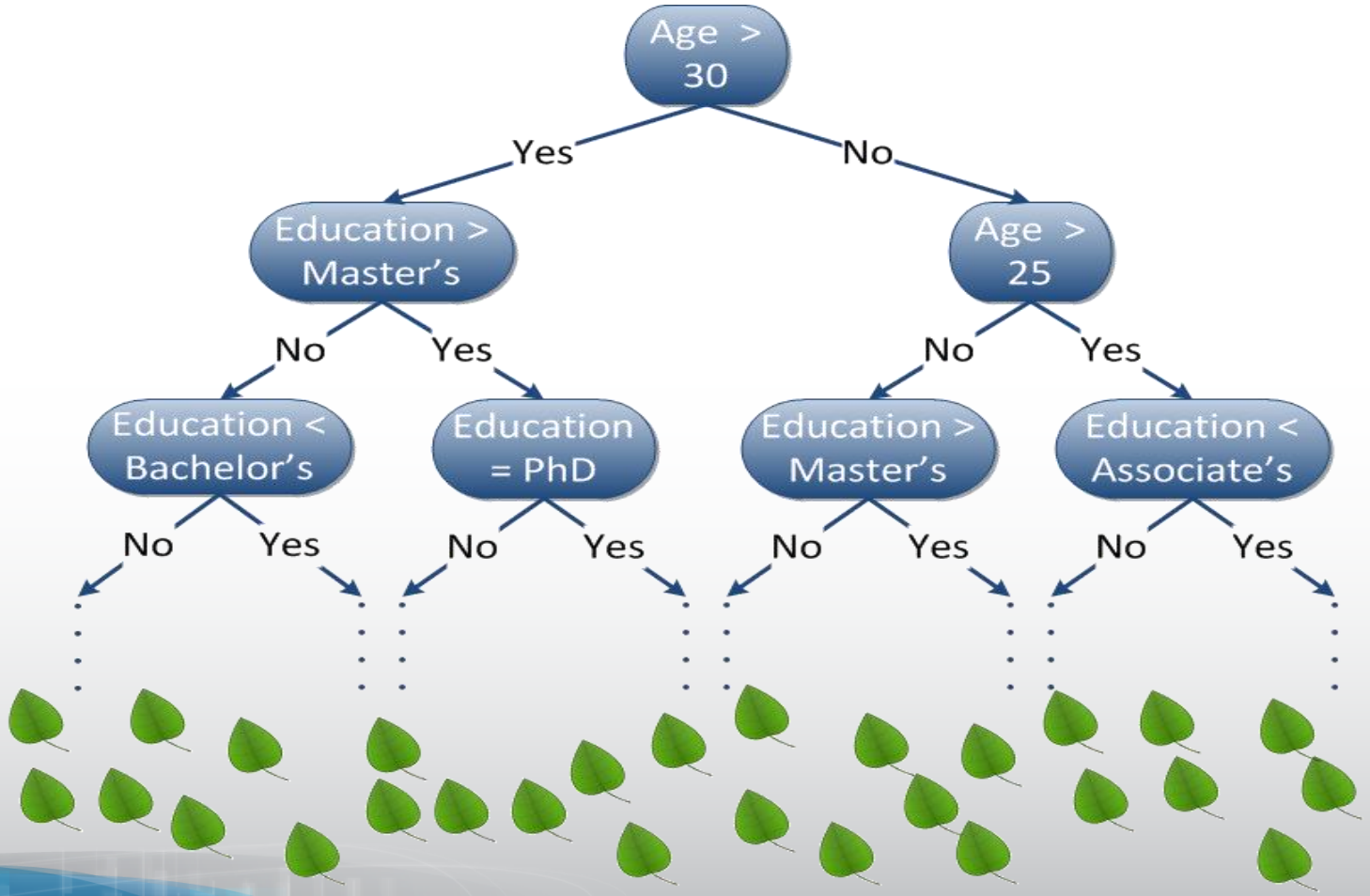
Classification and Regression Trees (CART)

- Estimates univariate outcomes conditional on multivariate predictors
- Produces recursive binary splits of the predictors to form relatively homogeneous groups
- Creates synthetic data by drawing values from "leaves"

The Synthesis Models

- Sex: Dirichlet model instead of CART
- Age | Sex
- Race | Age, Sex
- Hispanic Origin | Race, Age, Sex
- Educational Attainment | Hispanic Origin, Race, Age, Sex
- Marital Status | Educational Attainment , Hispanic Origin, Race, Age, Sex
- Wages | Marital Status , Educational Attainment , Hispanic Origin, Race, Age, Sex

Tree to Predict Wages



CART versus Parametric Models: Pros

- More easily applied, especially with irregular distributions
- Can capture non-linear relationships and interaction effects that may not be easily revealed
- Provides a semi-automatic way to fit the most important relationships in the data

-- Reiter, *Journal of Official Statistics*, 2005

CART versus Parametric Models: Cons

- Discontinuity at partition boundaries
- Decreased effectiveness when relationships can be accurately described by parametric models

-- Reiter, *Journal of Official Statistics*, 2005

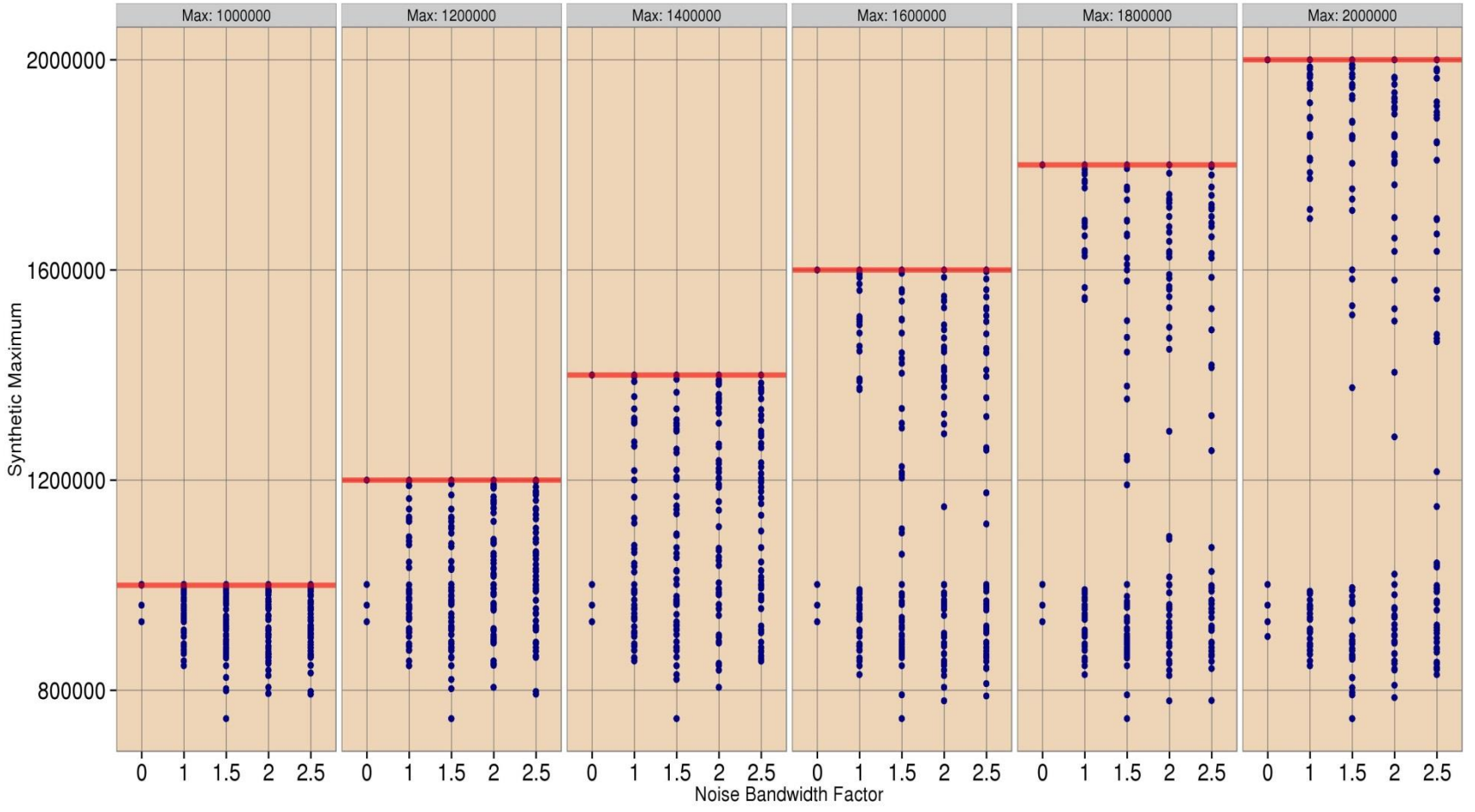
Concerns with Continuous Variables

- CART replaces data with actual observed values; could be too risky for continuous variables
- Proposed solution: Apply kernel density smoothing
  - First proposed by Reiter, Journal of Official Statistics, 2005
  - Included in R package “synthpop”
  - Implemented in some public datasets
- Disclosure Concerns
  - Attribute disclosure: Exact or within a range
  - Outliers are at particular risk
  - Person’s presence within a survey

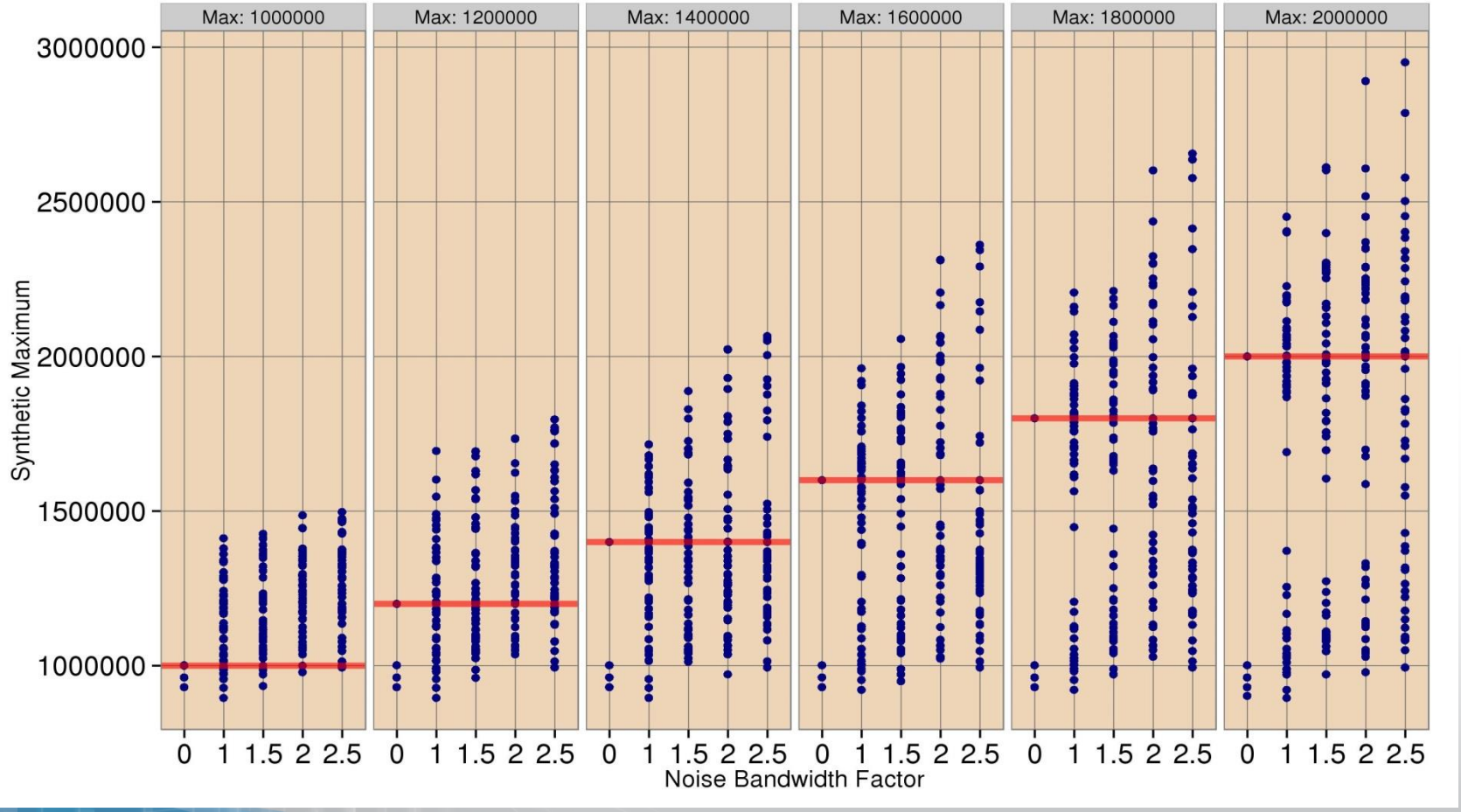
Density Evaluation

- Vary noise specifications on wages
  - Method 1: Kernel density support from bottom to the top of leaf
  - Method 2: Kernel density support extending above top of leaf for higher incomes
  - Both methods used for a variety of bandwidth sizes
- Disclosure Risk: Evaluate threat for original dataset's max income, for various maxes
- Data Utility: Compare mean wages

Synthetic Maxes for Method 1



Synthetic Maxes for Method 2

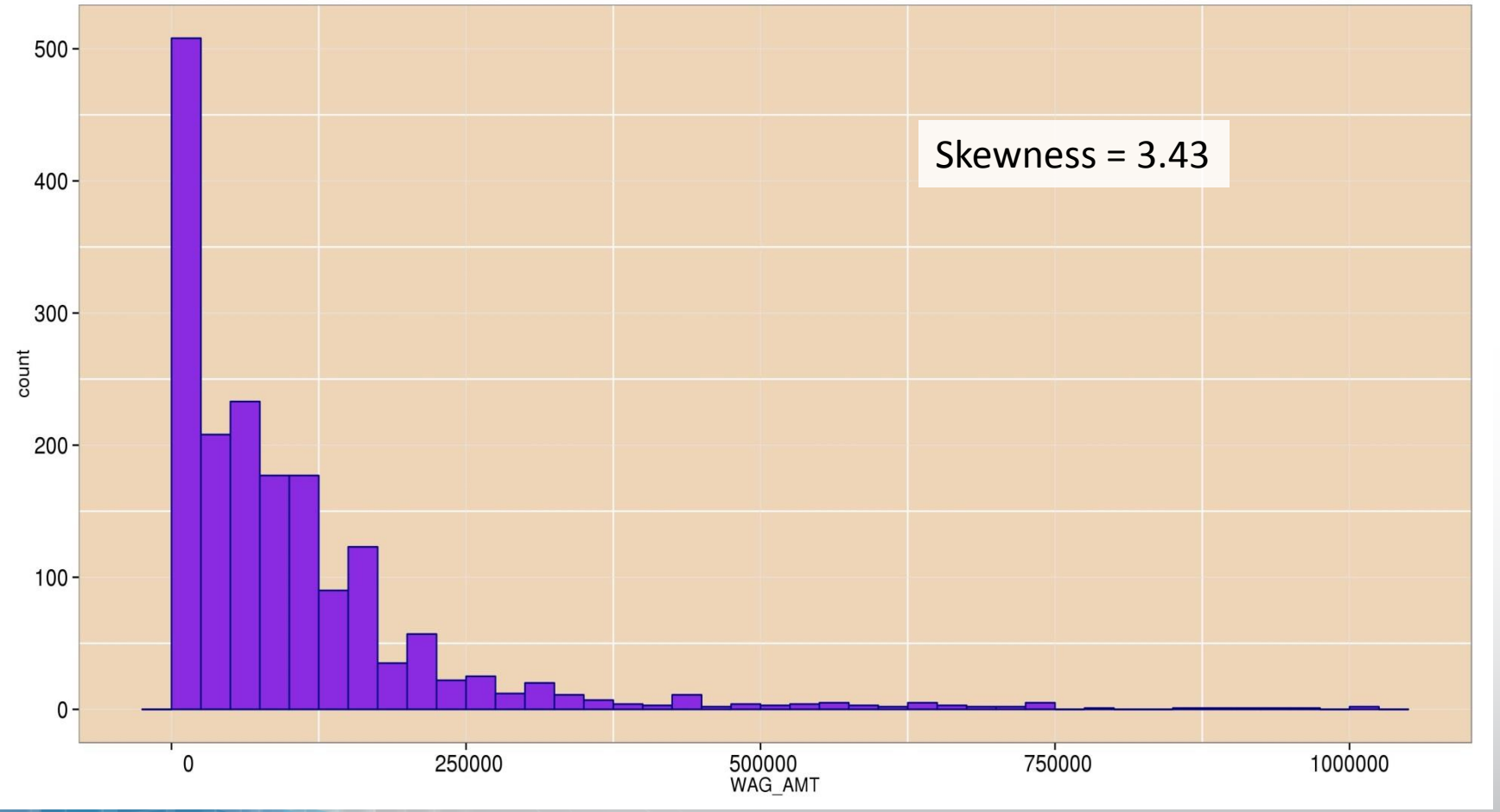


Mean Wages

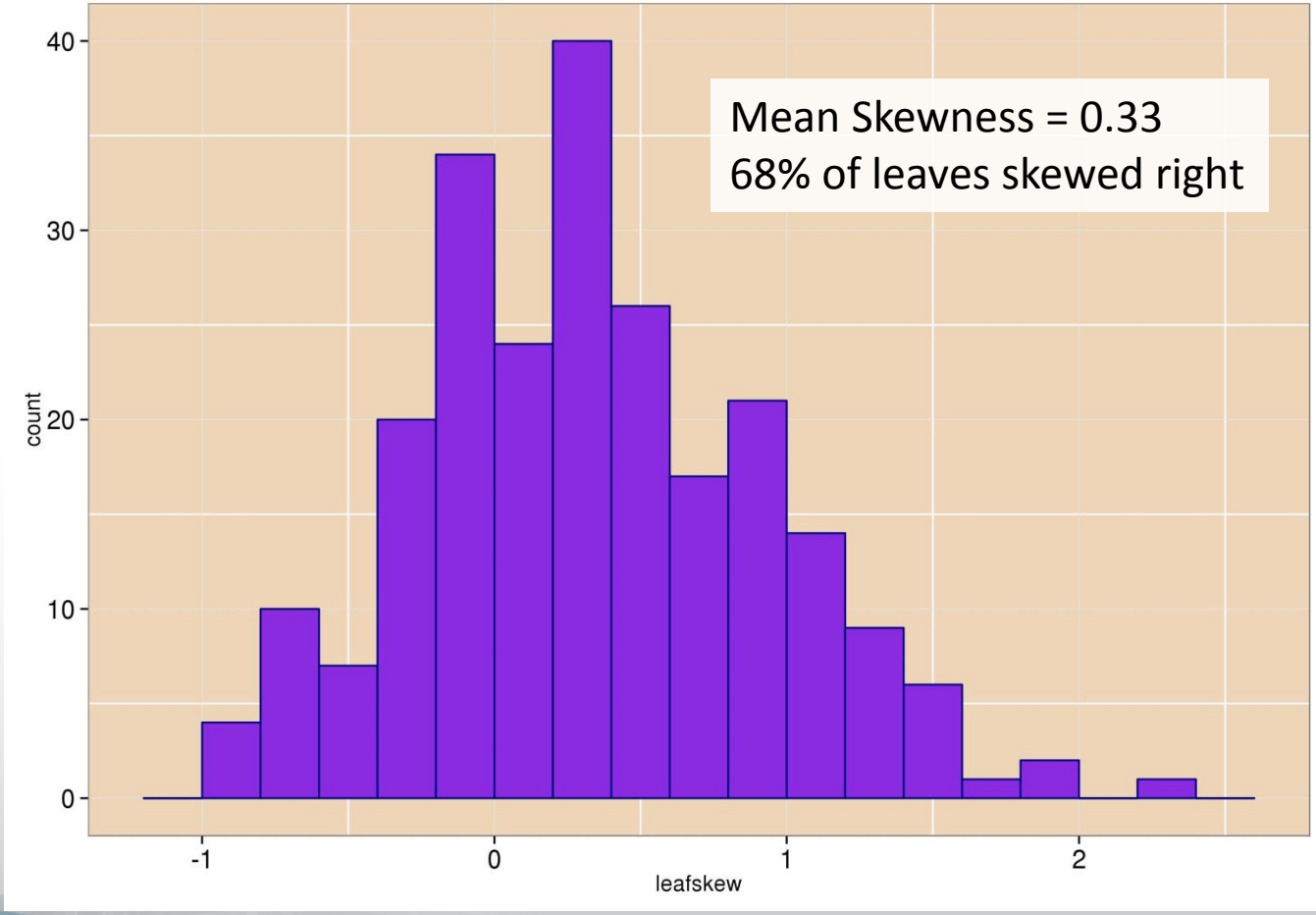
	Method 1		Method 2	
	Mean	SE	Mean	SE
Original	69,311		69,311	
Noise Factor = 1	69,210	2,088	73,521	2,405
Noise Factor = 2	71,482	2,147	**77,186**	2,579
Noise Factor = 3	73,647	2,236	**80,461**	2,753
Noise Factor = 4	**74,845**	2,273	**83,280**	2,892

**Estimates in bold are significantly different from the original mean.

Wages: Right Skewed



Leaf Skewness



Conclusions

- When releasing a single implicate, smoothing may provide enough protection if:
  - Bandwidth is large enough
  - Density support extends beyond leaf
- Attacks are possible if releasing multiple implicates
  - Method 1: Examine max of the synthetic maxes
  - Method 2: Examine average of the synthetic maxes
- Smoothing can potentially create biased estimates
- Some outliers may still be at risk

Next Steps

- Explore noise more
  - How to avoid bias?
  - Vary noise bandwidth by local dispersion?
  - Add noise before synthesis for risky values?
  - Topcode certain variables?
- Continue working on overarching questions

Questions?

Amy Lauger

Amy.d.lauger@census.gov

Thanks!