An Examination of the Effect of Differential Privacy Techniques on the Data Utility of Estimates of the Cost of Childcare

Quentin Brummet
NORC at the University of Chicago

Health Survey Research Methods Conference
March 6, 2020

NORC at the UNIVERSITY of CHICAGO

Motivation

- Increasing amounts of publicly retrievable data and advances in computational power have led to increased risk of reidentification
- Differential Privacy (DP) provides a suite of methods for adding a controlled amount of noise to estimates in order to protect respondent privacy
- As with any disclosure avoidance mechanism, this will result in a loss of utility
  - How this occurs will be context dependent

Differential Privacy

- A mathematical framework for protecting privacy
- The amount of privacy is governed by a privacy budget, typically denoted ϵ
- Two different approaches:
  - Add noise to each statistics separately → each additional statistic created will "spend" privacy budget
  - Create an underlying DP data set and calculate statistics from this → only have to "spend" privacy budget on creation of this data set
- ϵ is a measure of privacy, not data utility → different methods can produce much different results in terms of data utility while conforming the same level of privacy protection

What we do

- Perform comparisons of data utility from a variety of DP methods
- Use measures of cost of early care and education (ECE) from the National Survey of Early Care and Education (NSECE) as a test case
- Two different types of statistics:
  1. Mean summary statistics
  2. Regression coefficients from sensitive underlying data

Results

- While a given value of ϵ provides the same privacy (as measured by the DP privacy definition), data utility can differ when using different methods
- Simple methods that infuse noise to each estimate work well for a small number of estimates
- When releasing a large number of estimates, more complex methods are needed so that the privacy budget is used effectively
- Small decisions can have practically important effects
  - Examples: Number of bins in a histogram, range and standardization of variables

The National Survey of Early Care and Education

- Department of Health and Human Services-funded data collection
  - Household survey is of 11,629 U.S. households with at least one resident child under age 13 years
- NSECE makes sensitive data available through a variety of restricted access mechanisms
- Primary variable of interest: weekly costs for regular ECE for all children in a household up to age 13
  - May differ from common definitions within the existing literature and discussions regarding ECE costs

Mean Summary Statistics

- Consider Two Approaches:
  1. Laplace Mechanism
     - Basic mechanism that adds noise to the end estimate, where noise is calibrated to a measure of privacy loss
  2. Perturbed Histogram
     - Create a “histogram” of the data set that has noise infused and calculate statistics from this

- This does not consider potentially more advanced methods that would build on a perturbed histogram (e.g., the TopDown algorithm being employed for the 2020 Census)

Approach 1: Laplace Mechanism

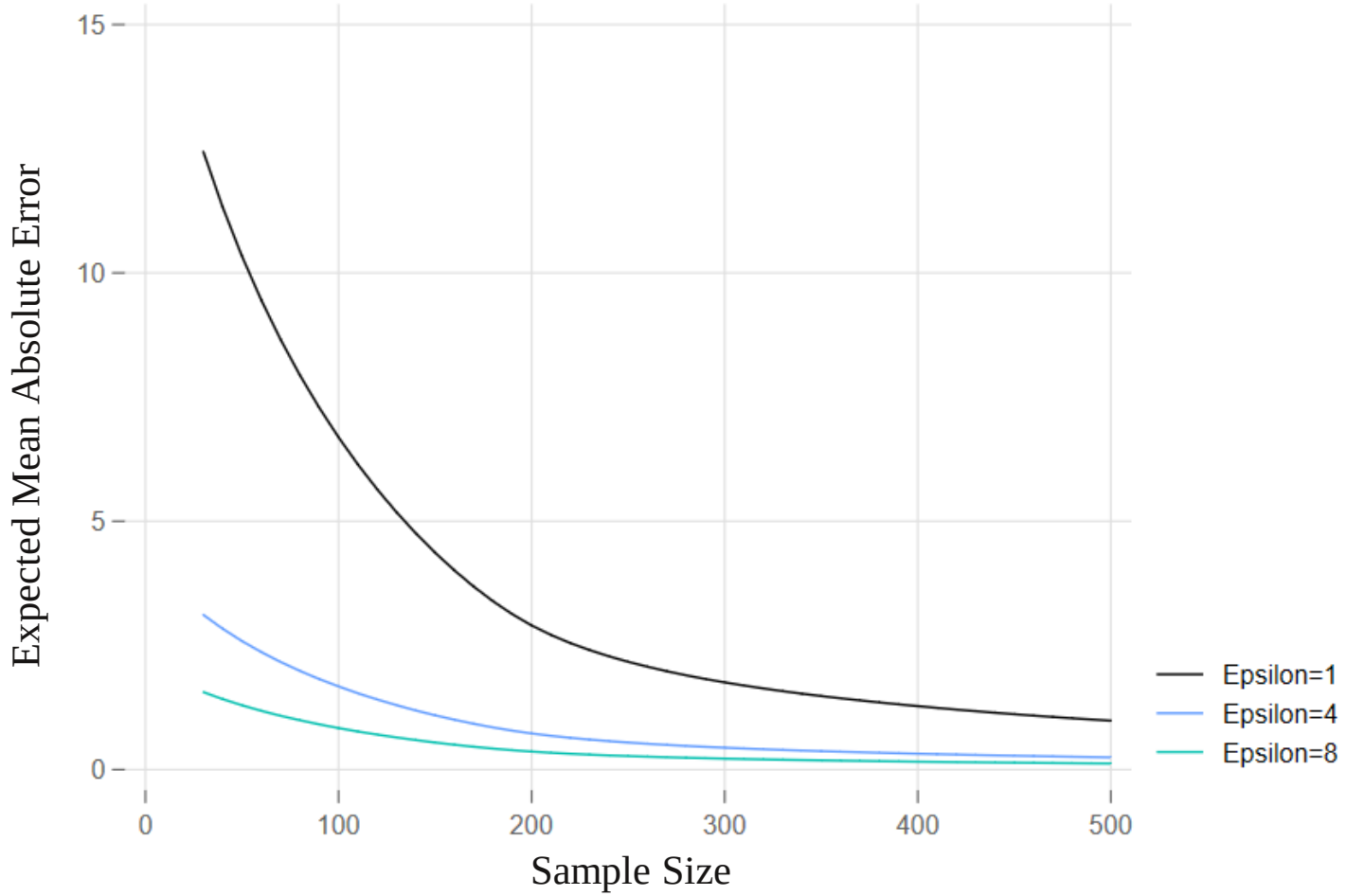
1. Calculate a statistic of interest

2. Add random noise to the estimate
   - Noise is calibrated based on the "sensitivity" of the estimate and the privacy budget, epsilon

In the context of a mean, noise follows the following distribution, where Λ is the range of the variable and n is the number of observations:

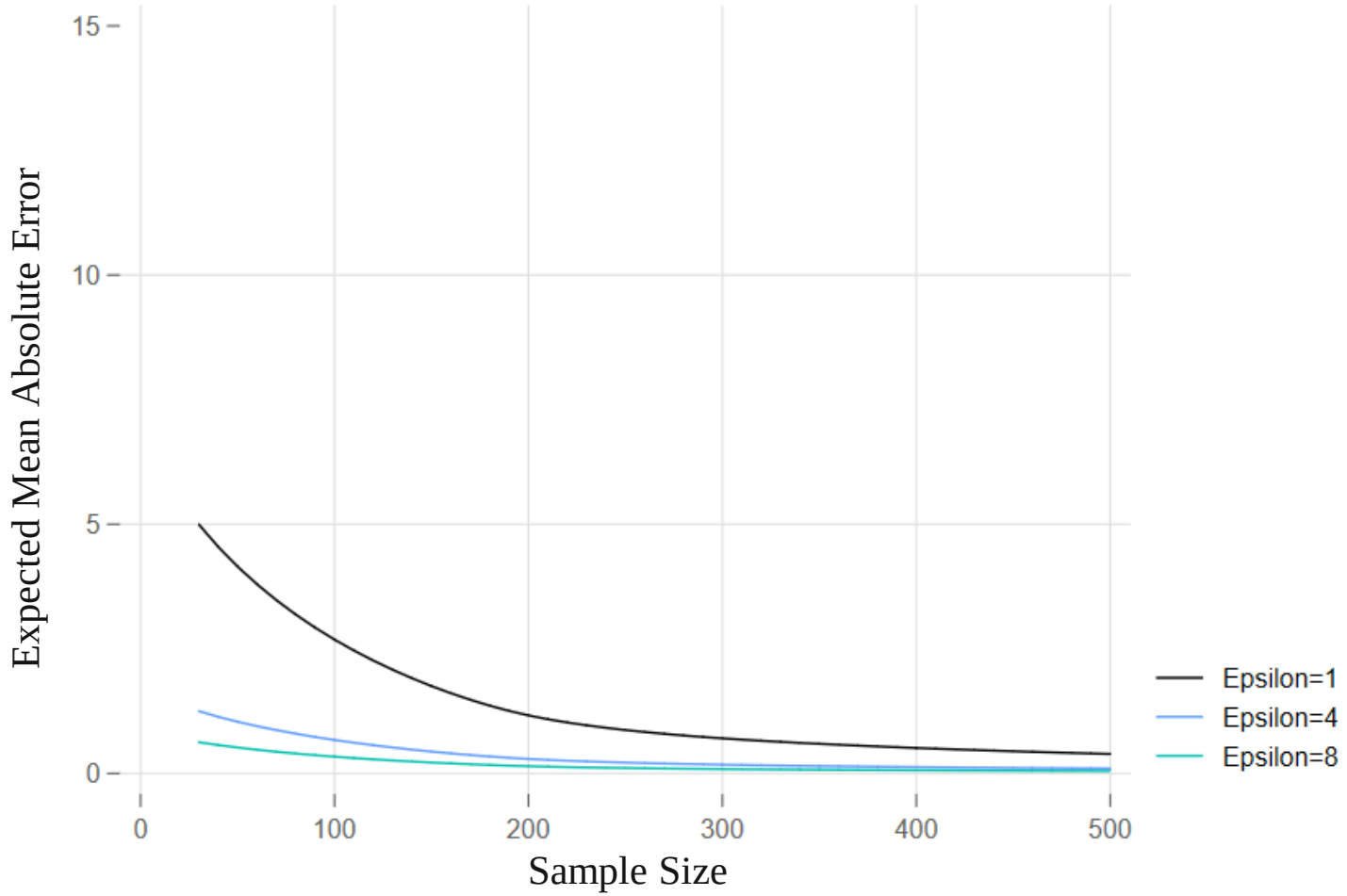
$$Laplace(0, \frac{\Lambda}{\epsilon n})$$

Error Induced by Laplace Mechanism, Variable Takes Values between 0 and 500 (i.e., $\Lambda = 500$)



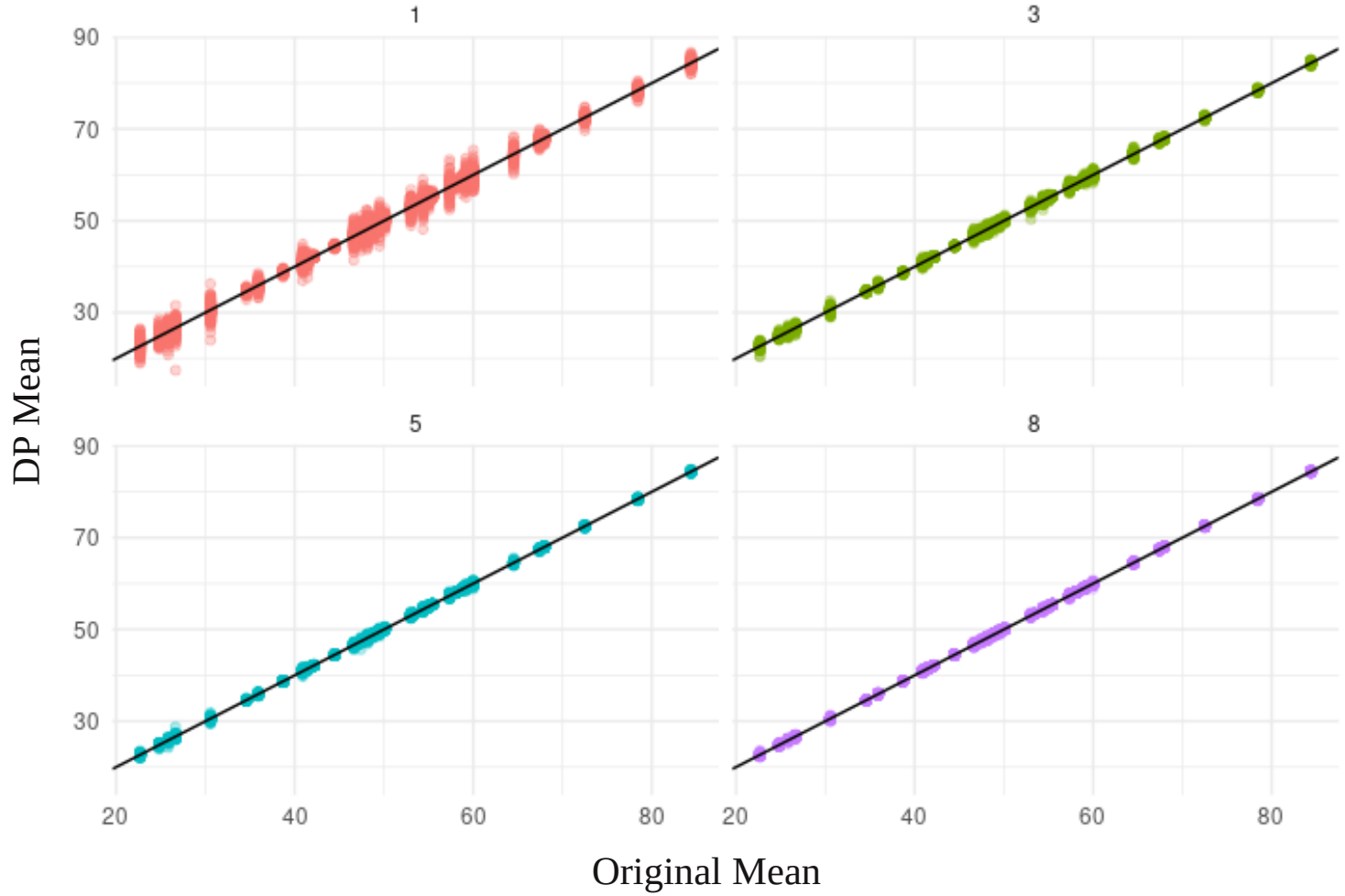
Source: Authors calculations based on simulated data.

Error Induced by Laplace Mechanism, Variable Takes Values between 0 and 200 (i.e., $\Lambda = 500$)



Source: Authors calculations based on simulated data.

Error across States using Laplace Mechanism

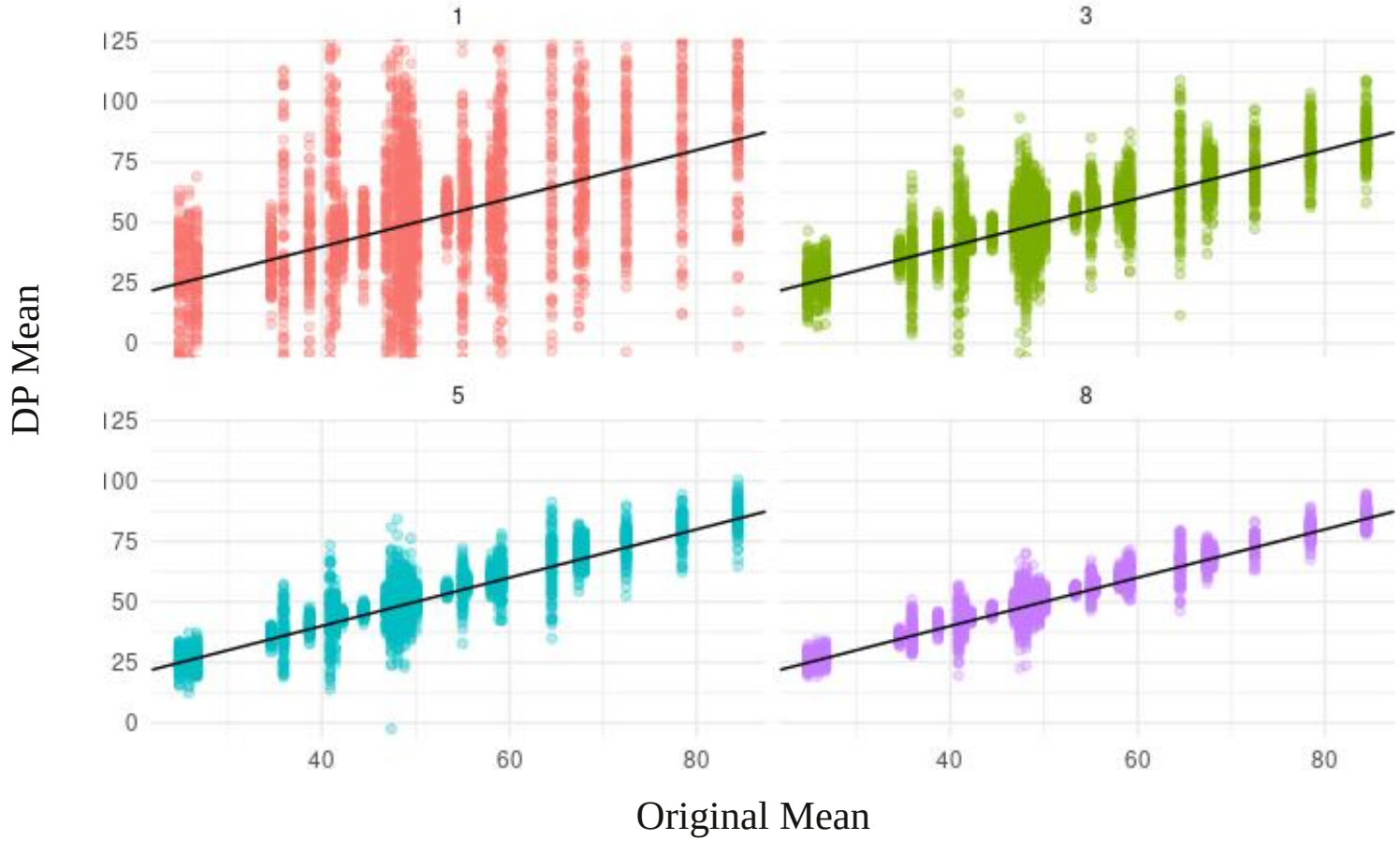


Source: 2012 NSECE Household Survey. Variable of interest is average weekly cost for all children up to age 13.

Approach 2: Perturbed Histogram

1. Divide the data into equally spaced bins
   - This must be done without looking at the private data
2. Collapse the data into counts in each bin
3. Add noise to each count separately
4. Calculate statistics based on this “perturbed histogram”

Error across States using Perturbed Histogram



Source: 2012 NSECE Household Survey. Variable of interest is average weekly cost for all children up to age 13. Y axis truncated at 0 and 120 for interpretability. Number of bins=100.

Perturbed Histogram by Bin Size

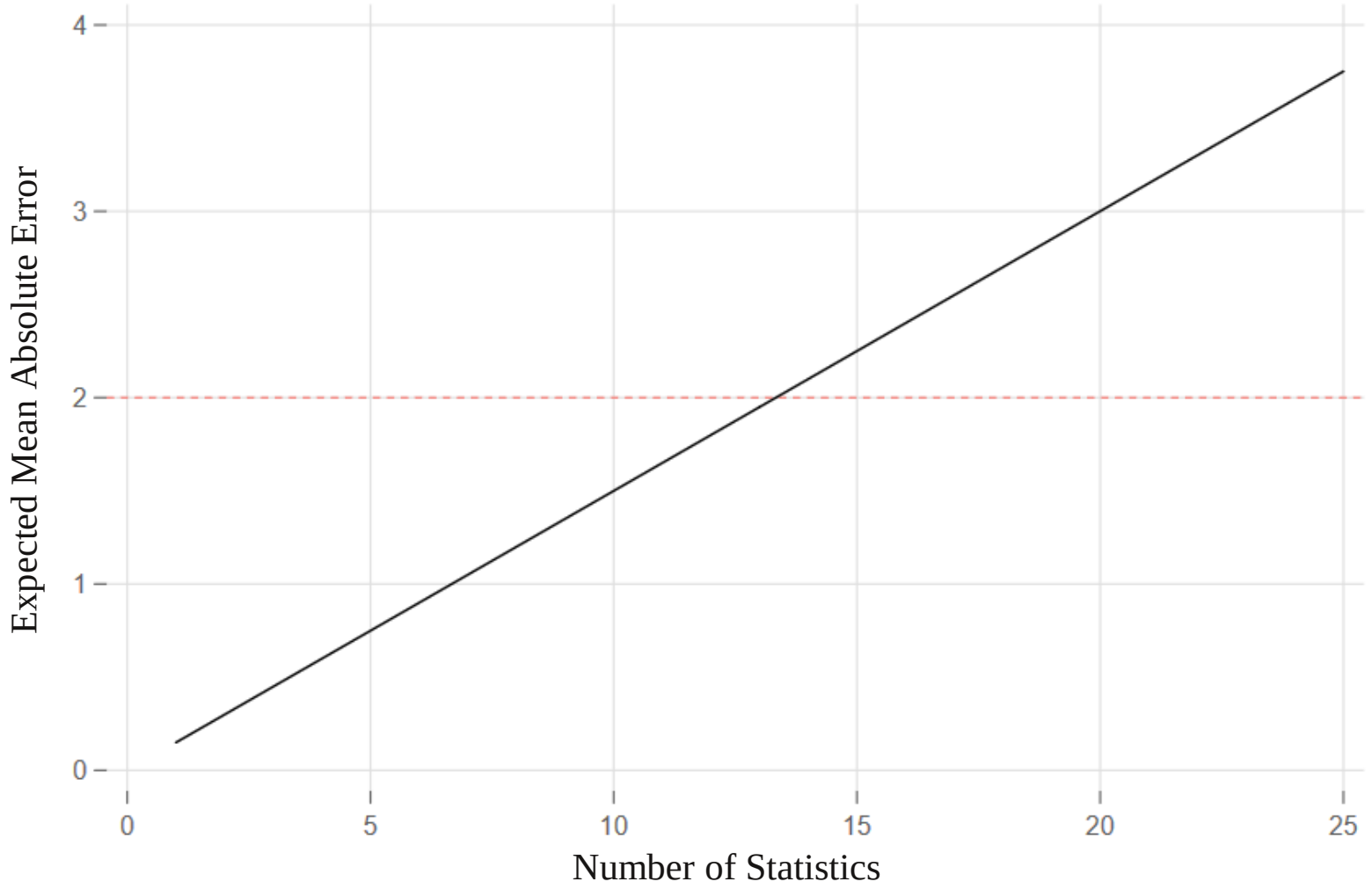
State	N	Mean	SE of Mean	Mean Absolute Error	
				50 Bins	100 Bins
CA	>=200	41.5	2.00	3.72	2.91
TX	>=200	53.4	3.19	4.76	5.46
NV	80-199	25.8	4.85	11.8	16.9
OK	80-199	26.7	5.10	10.9	14.4

Source: 2012 NSECE Household Survey. Mean figures refer to mean weekly cost of ECE for all households.

Comparison of Laplace and Perturbed Histogram Methods

- Hypothetical exercise:
  - Total privacy budget = 8
  - Take the state of ~500 sample size
- Based on prior results, we might expect the absolute error from a perturbed histogram approach to be ~$2/day
- Expected error from a Laplace mechanism would be dependent on the number of statistics released
  - Each additional statistic consumes more privacy budget

Expected Absolute Error by Number of Statistics Produced



Laplace Mechanism

Perturbed Histogram

Source: Authors calculations from prior results.

NORC at the UNIVERSITY of CHICAGO

Protecting Regression Coefficients Using DP

- Often, data providers wish to provide access to sensitive data in secure research environments and allow the disclosure of aggregate output that has passed through a disclosure review
  - One common example: regression coefficients
- There exist a number of DP techniques for regression coefficients, but they often differ according to specific types of regressions
- We will consider a simple regression of weekly costs on the ages of children in the household, the presence of family members nearby, and a measure of income relative to the poverty line
  - We consider unstandardized variables, though many DP implementations often standardize all variables prior to the regression

Methods for Adding DP Noise to Regression Coefficients

1. Sufficient Statistic Perturbation (SSP)
   - Adds Gaussian noise to the matrices that are used to create linear regression coefficients
   - Dwork et al. (2014)

2. Adaptive Sufficient Statistic Perturbation (AdaSSP)
   - Uses ridge regularization to mitigate additional error caused by outliers and collinear covariates
   - Wang (2018)

3. Objective Perturbation
   - For non-linear regressions, adds a penalty and random cross-product term to likelihood being maximized
   - Chadhuri et al. (2011), Kifer et al. (2012)

OLS Results: Dependent Variable = Weekly Cost, $\epsilon = 8$

	OLS	SSP (median)	AdaSSP (median)
N children age 0-5	23.4	0.50	17.6
N children age 6-8	13.6	-1.00	10.3
N children age 9-12	8.03	0.09	3.76
N children age 13-18	-3.58	0.07	-2.85
Relatives Nearby	-12.8	-0.18	-6.51
Income/Poverty Ratio	12.8	0.73	12.1

Source: 2012 NSECE Household Survey. Dependent variable is total weekly cost for ECE. SSP and AdaSSP columns report median coefficient estimate across 200 runs of the DP algorithm.

OLS Results: Logged Dependent Variable, $\epsilon = 8$

	OLS	SSP (median)	AdaSSP (median)
N children age 0-5	0.11	0.10	0.77
N children age 6-8	0.06	0.05	0.39
N children age 9-12	0.03	-0.01	0.52
N children age 13-18	-0.02	-0.05	0.27
Relatives Nearby	-0.07	-0.09	0.50
Income/Poverty Ratio	0.06	0.05	0.30

Source: 2012 NSECE Household Survey. Dependent variable is log(total weekly cost for ECE + 100). SSP and AdaSSP columns report median coefficient estimate across 200 runs of the DP algorithm.

DP Logistic Regression with Objective Perturbation, $\epsilon = 8$

	Nonprivate	ObjPert (median)
N children age 0-5	0.36	0.24
N children age 6-8	0.36	0.17
N children age 9-12	0.03	2.37
N children age 13-18	-0.12	0.16
Relatives Nearby	-0.40	0.07
Income/Poverty Ratio	0.31	0.10

Source: 2012 NSECE Household Survey. Dependent variable is an indicator for total weekly cost being greater than zero. ObjPert column reports median coefficient estimate across 200 runs of the DP algorithm.

Conclusion

- Simple methods work well for simple data releases, but applying DP is far from easy in most realistic settings
  - Typically requires using the privacy budget efficiently across multiple statistics
- Because there is no “one size fits all” approach, using DP methods effectively requires deep understanding of the subject matter
- Seemingly small decisions can have practically important effects
  - Example: regression coefficients can be especially sensitive to the scale of variables that are included in the model

Brummet-Quentin@norc.org

Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™

Simple Hypothetical Example

State	College Graduate?
Virginia	Yes
Virginia	No
Virginia	Yes
…	….
DC	Yes
DC	
…	….
Maryland	No
Maryland	No
…	….

	College Graduate?	
State	No	Yes
Virginia	18	15
DC	13	22
Maryland	16	16

Total N = 100.

Simple Hypothetical Example

- Draw noise from the Laplace distribution with scale parameter = $1/\epsilon$

	College Graduate?		
State	No	Yes	**Total**
Virginia	18	15	33
DC	13	22	35
Maryland	16	16	32
Total	47	53	100

	College Graduate?		
State	No	Yes	**Total**
Virginia	18 +1 = 19	15 – 3 = 12	31
DC	13 – 14 = -1	22 + 12 = 34	33
Maryland	16 – 2 = 14	16 + 7 = 23	37
Total	32	69	101