

An Initial Look At Variance in 2020 Census DAS

Jim Livsey*

Eric Slud[†]

September 3, 2020

Abstract

This is a working document purposed to investigate the variance in runs of the 2020 differentially private disclosure avoidance system.

Disclaimer This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

*Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233-9100, james.a.livsey@census.gov

[†]eric.v.slud@census.gov

1 Introduction and Statement of Problem

An important first step is to clearly state the problem, or set of problems, we want to answer. The 2020 DAS will be a privacy mechanism that acts upon the Census edited file (CEF).

$$\text{CEF} \longrightarrow \boxed{\text{DP}} \longrightarrow X^{\text{DP}} \longrightarrow \boxed{\text{Post-processing}} \longrightarrow \text{released data} \quad (1)$$

We can think of the CEF and differentially private dataset X^{DP} as a large dimensional histogram (contingency table). The DP box of (1) will add Laplace noise to specific entries of the contingency table. Assume our Laplace density is given by

$$\text{Lap}(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}}.$$

The value for the scale parameter b will change depending on the privacy budget ϵ . Then for different levels of geography and different table marginals the scale parameter will be:

$$b = G * Q * \epsilon$$

where G =geography level budget proportion and Q = query budget proportion OR detailed budget proportion. An example might be given by Table 1.

Table 1: An example of a possible allocation of the privacy budget

		Geographic Levels						
		National	State	County	Tract	BlockGroup	Block	Total
Queries	detailed	0.02	0.02	0.015	0.015	0.015	0.015	0.1
	hhgq	0.04	0.04	0.03	0.03	0.03	0.03	0.2
	votingage * hispanic * cenrace	0.1	0.1	0.075	0.075	0.075	0.075	0.5
	age * sex	0.04	0.04	0.03	0.03	0.03	0.03	0.2
	Total	0.2	0.2	0.15	0.15	0.15	0.15	1



1. Answer the question: How fast does the variance of a cell estimate decrease or level-off when adding additional marginal information. This question is important if developing an app, based on DP data, to answer a given query. If we find variance levels off with less the full number of marginals then this means we can answer a query with less then the full dataset. Hence, saving time and computing resources.



2. BLUE of true count X_{ij} is just the released value X_{ij}^{DP} with no additional marginal information. Adding marginal information can improve the quality of my estimate but by how much and how much marginal information is necessary to see these gains?
3. We treat single entry of table as unknown parameter and maximize likelihood with respect to the Laplace distribution.

2 Initial Toy Example

To get our feet wet working on this type of problem we investigate the following problem. Suppose we have a 2x2 contingency table of values. We can think of two outcomes of male/female and hispanic/non-hispanic. Denote the true table X :

	NHis	His	
male	X_{11}	X_{12}	X_{1+}
female	X_{21}	X_{22}	X_{2+}
	X_{+1}	X_{+2}	X_{++}

Suppose further we add independent Laplace noise to each entry $X_{i,j}$ as well as each marginal. The Laplace distribution (centered at 0) with scale parameter b has pdf

$$\text{Lap}(x|b) = \frac{1}{2b} e^{-\frac{|x|}{b}}.$$

Let $\xi_{\ell,k}$ be independent Laplace random variables for $\ell, k \in \{1, 2, +\}$. Adding noise to all entries of the contingency table X will yield a differentially private dataset X^{DP} such that

$$X_{ij}^{\text{DP}} = X_{ij} + \xi_{ij}. \quad (2)$$

Assuming we are given X^{DP} , the implied system with unknowns $X_{11}, X_{12}, X_{21}, X_{22}$ to be minimized is:

$$|X_{11} - X_{11}^{\text{DP}}| + |X_{12} - X_{12}^{\text{DP}}| + |X_{21} - X_{21}^{\text{DP}}| + |X_{22} - X_{22}^{\text{DP}}| + \quad (3)$$

$$|X_{11} + X_{12} - X_{1+}^{\text{DP}}| + |X_{21} + X_{22} - X_{2+}^{\text{DP}}| + |X_{11} + X_{21} - X_{+1}^{\text{DP}}| + |X_{12} + X_{22} - X_{+2}^{\text{DP}}| + \quad (4)$$

$$|X_{11} + X_{12} + X_{21} + X_{22} - X_{++}^{\text{DP}}| \quad (5)$$

The different parts of this equation can be understood as (5) the middle cells of the contingency table compared directly to their private released value, how the cells add to their marginal values are given by (3) and the total absolute deviation is given by (5).

Suppose the CEF values for these cells were given as:

	NHis	His	
Male	15.00	4.00	19.00
Female	3.00	8.00	11.00
	18.00	12.00	30.00

Table 2: Fictitious True CEF values for toy example

It is of primary interest to solve (3) - (5), henceforth known as the L1 problem, for amount of marginal information provided. We replicate 5000 times noise infused from Laplace distribution with $b = 1/3$ and compare the solution to the L1 problem to the known truth given in Table 2.

To establish a baseline for comparison we solve the L1 problem given no marginal information and for all margins and totals. This is displayed in Table 3.

2.1 Hard constraints

To mimic the invariants that will be a part of the 2020 DAS, we investigate how to add marginal constraints to the L1 problem. For the toy problem this would manifest as a constraint of the form "total number of males in the DP table must be 19".

We assume the invariants are known and hence any marginal DP observation will be our fixed value. For sake of example, assume we want to hold the Male population invariant as X_M in the L1 problem. Then (3) - (5) could be reformulated as

$$|X_M - X_{12} - X_{11}^{\text{DP}}| + |X_{12} - X_{12}^{\text{DP}}| + |X_{21} - X_{21}^{\text{DP}}| + |X_{22} - X_{22}^{\text{DP}}| + \quad (6)$$

$$|0| + |X_{21} + X_{22} - X_{2+}^{\text{DP}}| + |X_M - X_{12} + X_{21} - X_{+1}^{\text{DP}}| + |X_{12} + X_{22} - X_{+2}^{\text{DP}}| + \quad (7)$$

$$|X_M - X_{12} + X_{12} + X_{21} + X_{22} - X_{++}^{\text{DP}}| \quad (8)$$

where we have explicitly solved for $X_{11} = X_M - X_{12}$.

	None	All
X_{11}	0.50	0.37
X_{21}	0.47	0.37
X_{12}	0.51	0.43
X_{22}	0.50	0.37



Table 3: Baseline for all other toy examples. The left column give RMSE when no marginal information is given to the L1 problem. The right column gives RMSE when all information is given.




	$X_{+1}X_{+2}$	$X_{+1}X_{1+}$	$X_{+1}X_{1+}X_{++}$	$X_{+1}X_{+2}X_{++}$	$X_{+1}X_{+2}X_{1+}$	$X_{+1}X_{+2}X_{1+}X_{2+}$	all
1	0.47	0.36	0.35	0.44	0.35	0.35	0.35
2	0.48	0.48	0.37	0.45	0.48	0.35	0.36
3	0.47	0.48	0.37	0.44	0.34	0.36	0.38
4	0.49	0.47	0.45	0.44	0.45	0.36	0.37

Table 4: RMSE of L1 problem for 5000 replications of Laplace noise with $b = 1/3$ for different amount of marginal information.

3 Simulated Schema

To investigate the ideas outlined in Section 1, we conduct a simulation study of a hypothetical DAS schema. The variables included, different levels of each variable and total number of cells in the full cross table is shown in Table 5. We apply the the budget allocation given in Table 1 to a total budget of $\epsilon = 4$.

We will assume the tract, county, state nesting structure follows Table 6.

Table 5: Add caption				
		levels	#	Full cross
Variables	HHGQ	HH, GQ	2	168
	Sex	M, F	2	
	CenRace	Wh, Bl, As, AIAN, Pac, Oth, 2+	7	
	Age	0-17, 18-62, 63-115	3	
	Hispanic	Hisp, NonHisp	2	
Geography	Tract		20	
	County		7	
	State		2	

Table 6: Geography hierarchy for simulation																				
State level	1										2									
County level	1		2		3					4		5			6			7		
Tract level	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

3.1 Results

We compare the variance of 3 different types of marginal information.

- 1. All margins - baseline.
- 2. No margins - worst case.
- 3. Any margin pertaining to certain cell.



The specific cell in situation 3. above will be

Table 7: Specific cell we focus on for simulation variance						
HHGQ	SEX	RACE	AGE	HISP	GEO	Cell true value
hh	male	white	0-17	hisp	1	2

4 Assumptions

Here we explicitly state the assumptions we will be making.

- 1. Laplace Noise Parameters are known.

Table 8: Corresponding marginals for specific cell

Query	Array index	True Value
HHGQ	1, 0, 0, 0, 0, 1	478
votingAge x Hisp x cenRace	0, 0, 1, 1, 1, 1	14
Age x Sex	0, 1, 0, 1, 0, 1	149

5 Vocabulary

We use this section to book keep words and definitions from the DP world. More importantly, this is our understanding of the terminology.



Invariant Set of marginals that will be strictly enforced when post processing. Example: the released state total population will match the CEF state total population.

Workload Set of cells in histogram that need more/less noise to be added independent of remaining cells. Resulting in more/less precision of released estimates. Additionally, these are the **only** margins that will have noise added directly to them. Any other margin will not be needed in the L1 problem.

Query histogram entry or marginal. The workload is a set of queries.

6 L1pack R package

This section documents how we implement the Barrodale and Roberts algorithm. L1 estimation for linear regression, density, distribution function, quantile function and random number generation for univariate and multivariate Laplace distribution.

Basic usage:

```
l1fit(x, y, intercept = TRUE, tolerance = 1e-07, print.it = TRUE)
```

Arguments:

x vector or matrix of explanatory variables. Each row corresponds to an observation and each column to a variable. The number of rows of x should equal the number of data values in y, and there should be fewer columns than rows. Missing values are not allowed.

y numeric vector containing the response. Missing values are not allowed.

7 Reference List

7.1 Papers to Read

- [4] Is plenary lecture from *Theory and Applications of Models of Computation*. 5th International Conference, TAMC 2008.
- [6] work dealing specifically with Laplace noise

- [1] Book on LAD
- [3] "Analysis of least absolute deviation"
- [2] Conference proceedings paper

7.2 Brief summary

- [5] Portnoy and Koenker Statist. Sci. 1997. They give great history of ℓ_1 -methods. Discuss computational issues with simplex for ℓ_1 problems and say these can be alleviated using an 'interior search' of the manifold coupled with some pre-processing steps.

References

- [1] Peter Bloomfield and William L Steiger. *Least absolute deviations: theory, applications, and algorithms*. Springer, 1983.
- [2] Anne-Sophie Charest. Empirical evaluation of statistical inference from differentially-private contingency tables. In *International Conference on Privacy in Statistical Databases*, pages 257–272. Springer, 2012.
- [3] Kani Chen, Zhiliang Ying, Hong Zhang, and Lincheng Zhao. Analysis of least absolute deviation. *Biometrika*, 95(1):107–122, 2008.
- [4] Cynthia Dwork. Differential privacy. *Encyclopedia of Cryptography and Security*, pages 338–340, 2011.
- [5] Stephen Portnoy, Roger Koenker, et al. The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4):279–300, 1997.
- [6] Rathindra Sarathy and Krishnamurty Muralidhar. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Trans. Data Privacy*, 4(1):1–17, 2011.