OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Estimating the Variance of Complex Differentially Private Algorithms

Robert Ashmead

JSM 2019, Denver, Colorado

Collaborators

John Abowd, Philip Leclerc, and William Sexton of the U.S. Census Bureau and the entire team working on differentially private disclosure avoidance methods for the 2020 Decennial Census.

OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Disclaimer

This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not the U.S. Census Bureau.

Research Question

- Most common differentially private algorithms have known, closed-form variances that are not dependent on the true query answer itself.
- How do we estimate the variance for more complex methods which do not necessarily meet these properties?

Differential Privacy

Definition

A randomized algorithm M is ϵ-differentially private if for all $S \subset R$ and for all *neighboring* datasets x, y:

$$Pr[M(x) \in S] \leq e^{\epsilon} Pr[M(y) \in S]$$

Where R is the output space of M and the randomness is solely due to the algorithm M.

The Privacy-Loss Budget

One of the best features of differential privacy is the way one can track the (global) privacy-loss budget, ϵ, of a mechanism from its possibly many sub-components.

The privacy-loss budget can be translated into a worst-case bound on an attacker's ability to improve their inference about a person's data upon seeing the mechanism output relative to a counterfactual baseline of the inference the attacker would have made if that person's data had been deleted/changed/never collected before running the mechanism.

Common DP Algorithms Have Known Variance

The Laplace distribution (two-sided exponential distribution) (centered at 0) with scale b has pdf:

$$Lap(y|b) = \frac{1}{2b} e^{\left(-\frac{|y|}{b}\right)}$$

$$\text{Variance} = 2b^2$$

Given data x and a linear query f with sensitivity Δf , the Laplace Mechanism is defined as $M(x|\epsilon) = f(x) + Y$ where $Y \sim Lap(\Delta f/\epsilon)$

The variance of the Laplace mechanism is location invariant, meaning it doesn't depend on the value of $f(x)$.

Other Mechanisms Also Have Known Variance

- The (two-sided) Geometric Mechanism has variance

$$2 * \frac{e^{\frac{-\epsilon}{\Delta f}}}{(1 - e^{\frac{-\epsilon}{\Delta f}})^2}$$

- The matrix mechanism (Li, et al., 2015) used for answering many queries simultaneously based on a strategy matrix, also has known and location invariant variance

Post-Processing DP algorithms can improve accuracy, but complicates the variance

- Enforcing Non-negativity
- Maintaining Integers with (controlled) rounding
- Constraints to (known or invariant) marginals

Any post-processing is allowed as long as it only utilizes the output of the DP mechanism and not the input

Post-processing changes the properties of the variance.

The variance could depend on the true query answer which is not known.

A Simple Example

Apply the Laplace mechanism to a query answer with sensitivity 1 and with $\epsilon = 0.1, 1, 10$. Enforce non-negativity.

True query answer $= 1$

ϵ	Variance	Variance, non-negativity	Bias, non-negativity
0.1	200	79.37	4.53
1	2	1.25	0.17
10	0.02	0.02	0.0

True query answer $= 10$

ϵ	Variance	Variance, non-negativity	Bias, non-negativity
0.1	200	122.35	1.82
1	2	1.98	0.0
10	0.02	0.02	0.0

A More Complicated Example

In the "Topdown" algorithm for the Disclosure Avoidance System (DAS) for the 2020 Decennial Census the algorithm post-processes the differentially private estimates to enforce

- Non-negativity
- Integer answers
- Constraints to invariant marginals
- Hierarchical consistency between tables

Variance Estimation Options

- Use additional privacy-loss budget to estimate the difference between the released DP query estimates and the true estimates
- A rough approximation based on location-invariant closed form methods
- Monte Carlo methods
  - Can we just simulate the mechanism + post-processing?
    - Yes, but we would have to utilize additional privacy-loss budget
  - Proposed "Parametric Bootstrap" method

Proposed "Parametric Bootstrap" Method

Let d be our dataset, $M()$ be our DP mechanism, and $q()$ be a query of interest.

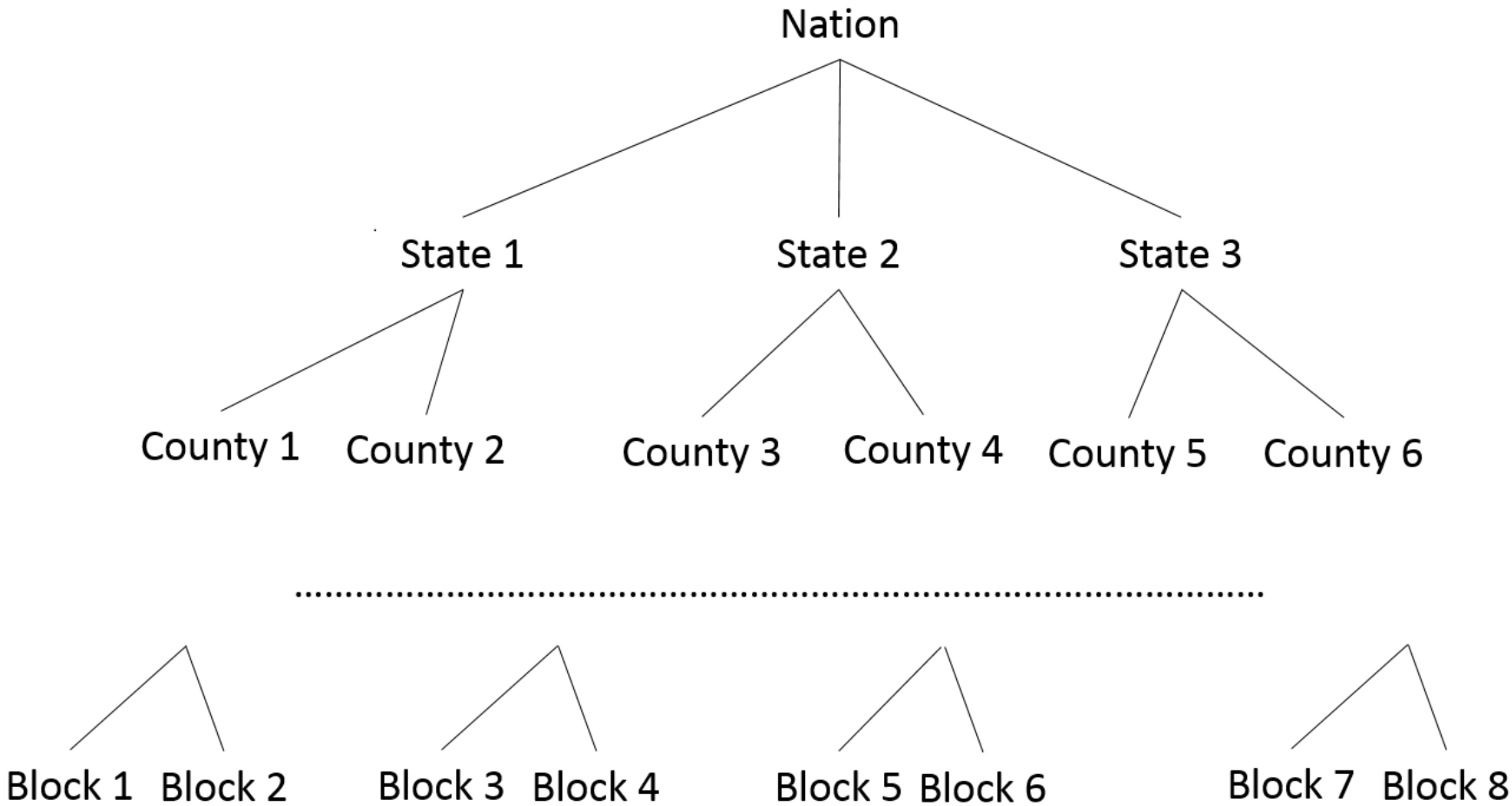
Suppose our mechanism releases an estimate of the dataset itself $\hat{d} = M(d)$.

If $\hat{d}$ is a reasonably accurate estimate of d, then might it be used to approximate the variance

$$Var(q(M(d))) \approx Var(q(M(\hat{d})))?$$

We do not need to spend the privacy-loss budget to simulate Monte Carlo draws of $q(M(\hat{d}))$

"Topdown" Mechanism as a Tree



"Topdown" Mechanism Summary

A. Take noisy histogram measurements using ϵ_1
B. Solve a constrained non-negative least-squares optimization problem which minimizes the squared distance between the solution and the noisy measurements, has a non-negative solution, and meets the constraints.
C. Solve a constrained rounding problem, which finds a nearby non-negative integer solution minimizing the distance from the LS solution (step B.) and also meets the constraints.
D. The solution is the privacy-protected histogram

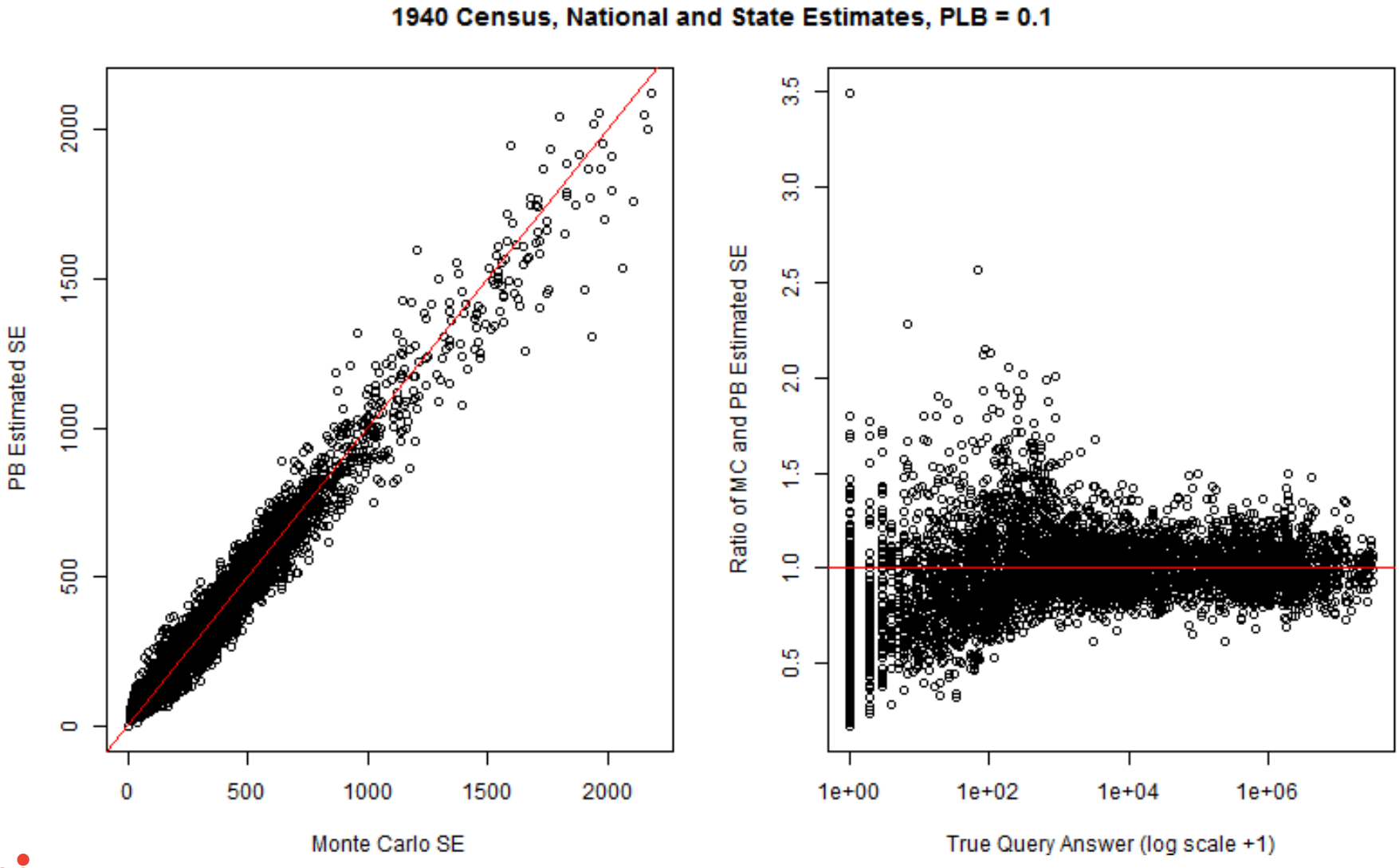
1940 Decennial Census Data Summary

- Data available from IPUMS (Ruggles et al., 2018)
- Geography levels (4): nation, state, county, enumeration district
- Schema: 8 x 2 x 5 x 5 x 6 = 2400 cells
- Variables: GQ/HH type, voting-age, Hispanic, citizen, race
- 132,404,766 total persons; 134,857 enumeration districts:
- 2400*134,857 = 323,656,800 total cells
  - Almost 3 times as many cells as total persons

Simulation Summary

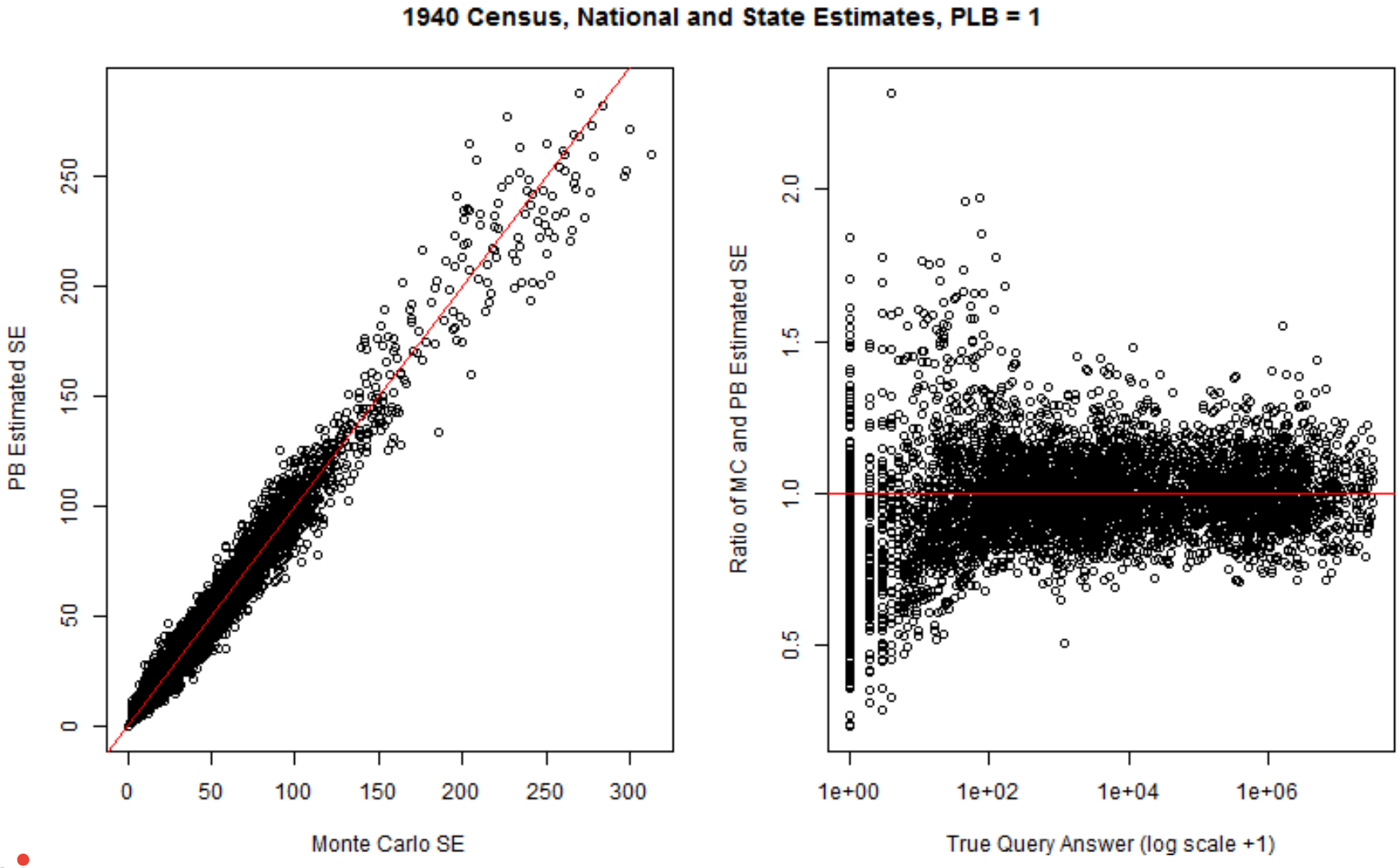
- For privacy-loss budgets of $0.1, 1.0$, and 5.0 estimate the variance of a number of queries at different geographic levels.
- Queries are a variety of marginal and crosses of the different variables
- Nation, State, and County
- Estimate the variance using both the Monte Carlo (MC) method (truth) and the proposed Parametric Bootstrap (PB) method.
  - The PB method uses the first run of the MC method as its estimate of the truth
  - Based on $n = 100$ simulations in both cases

Results



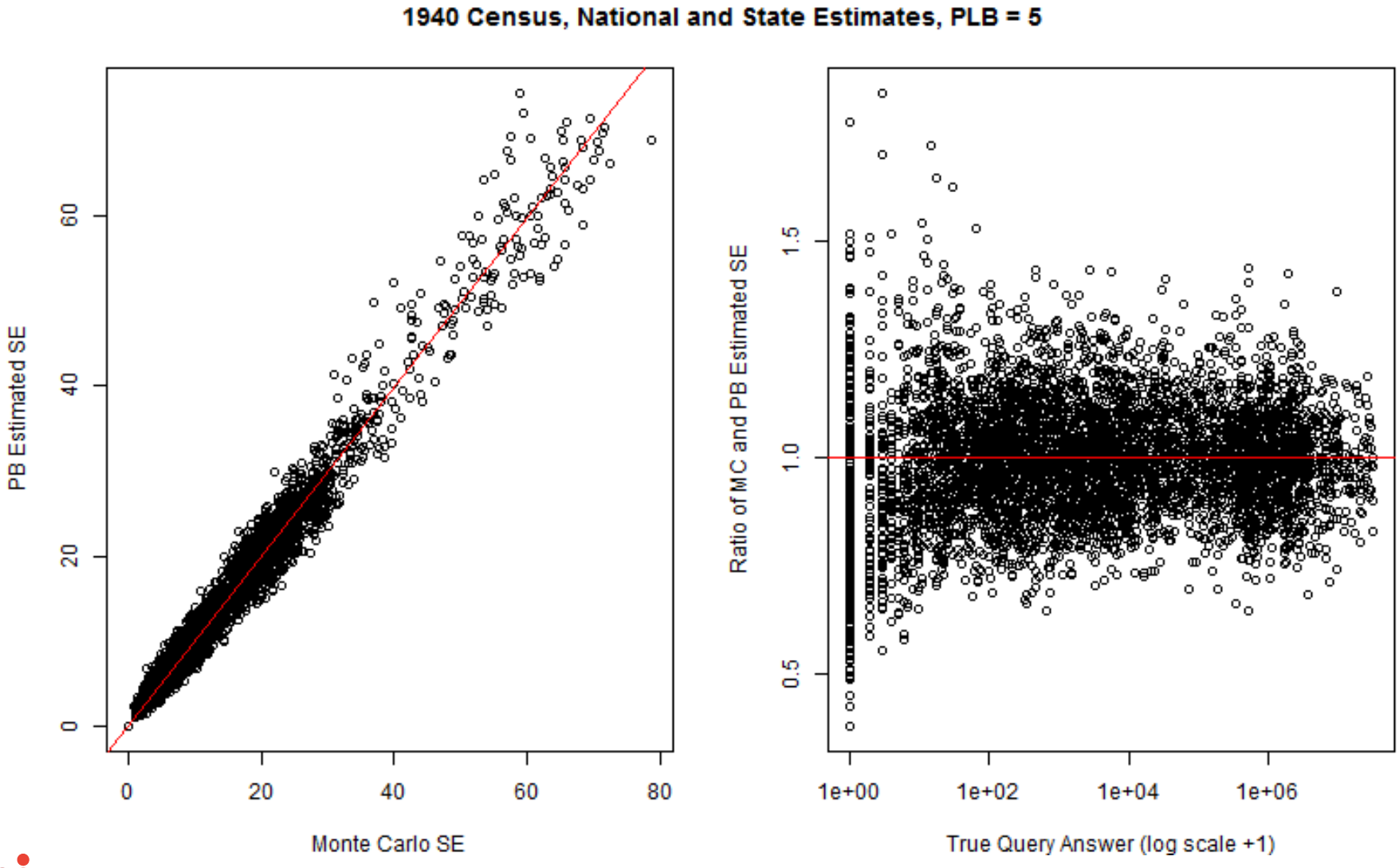
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results



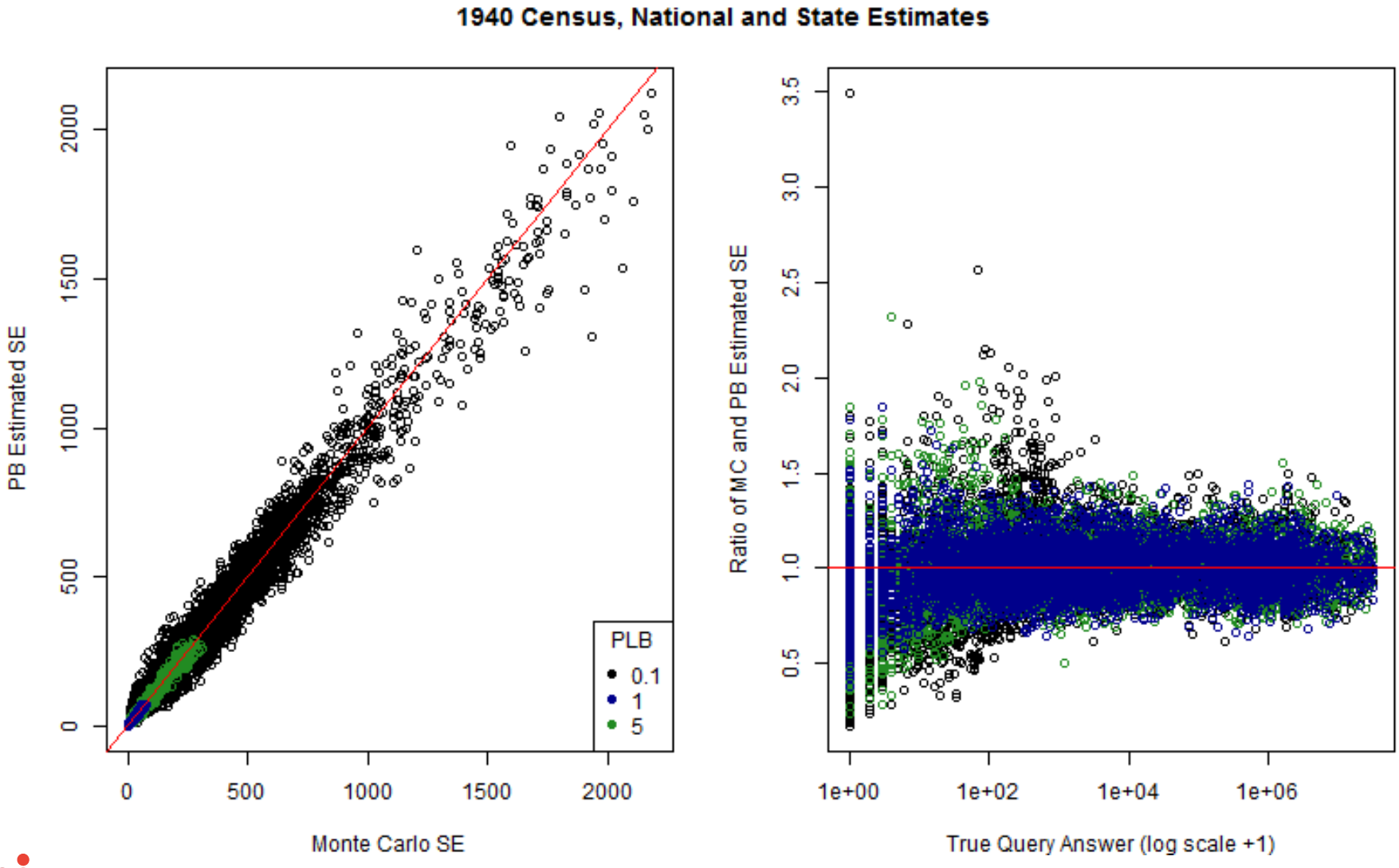
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results



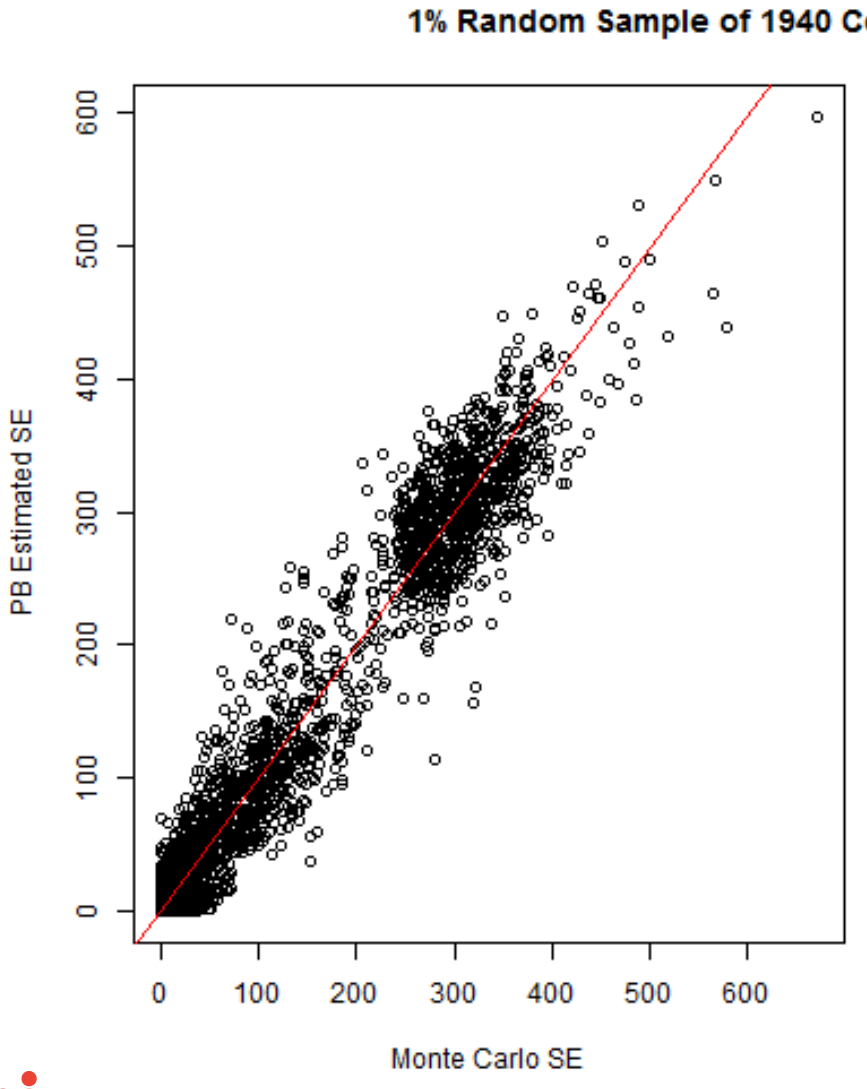
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

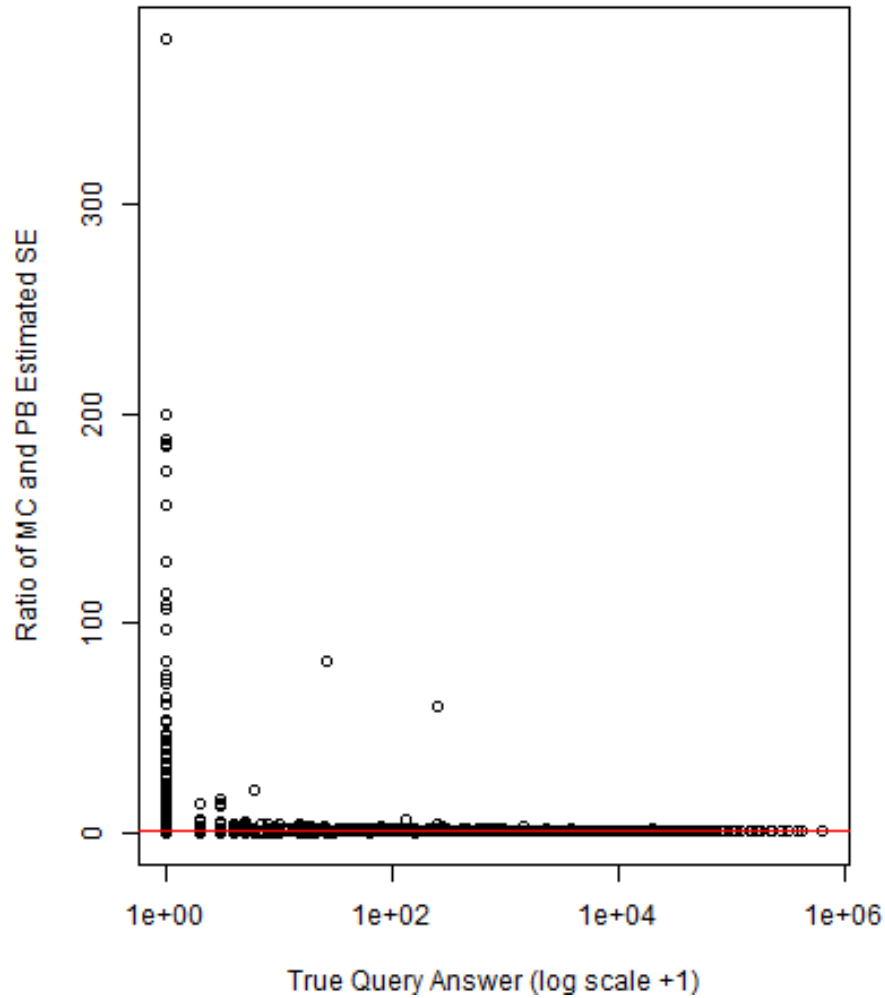
Results



OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

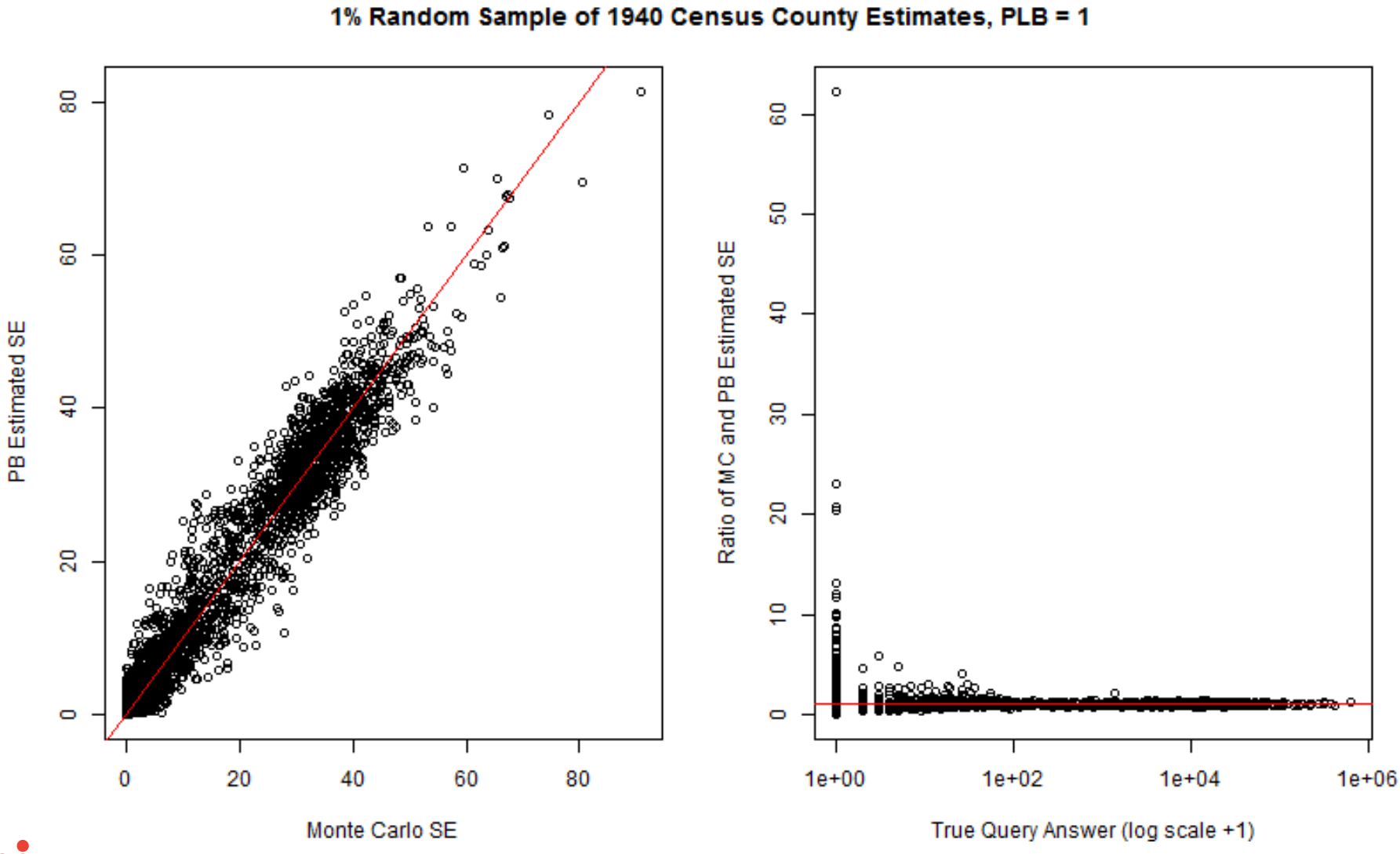
Results II





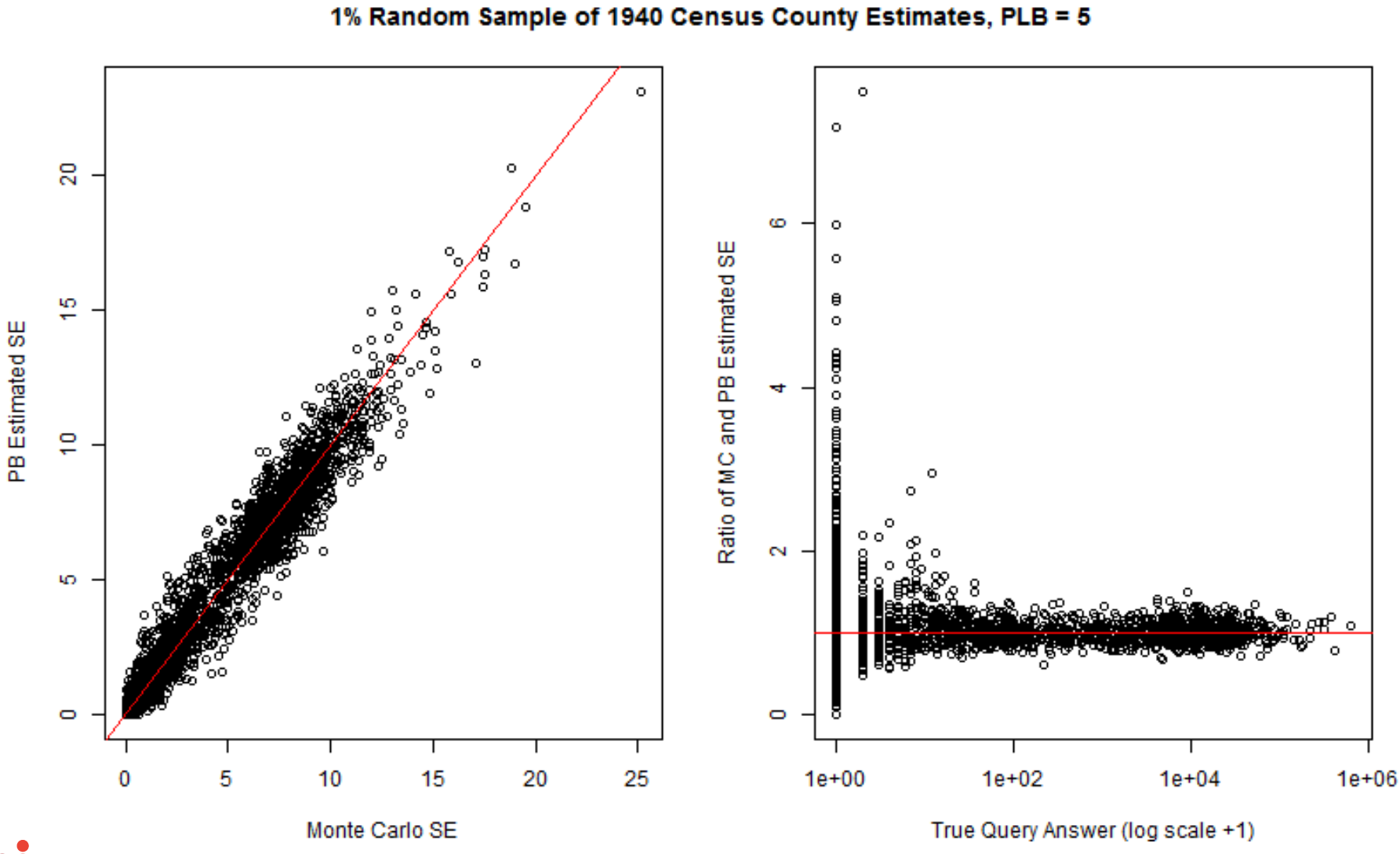
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results II



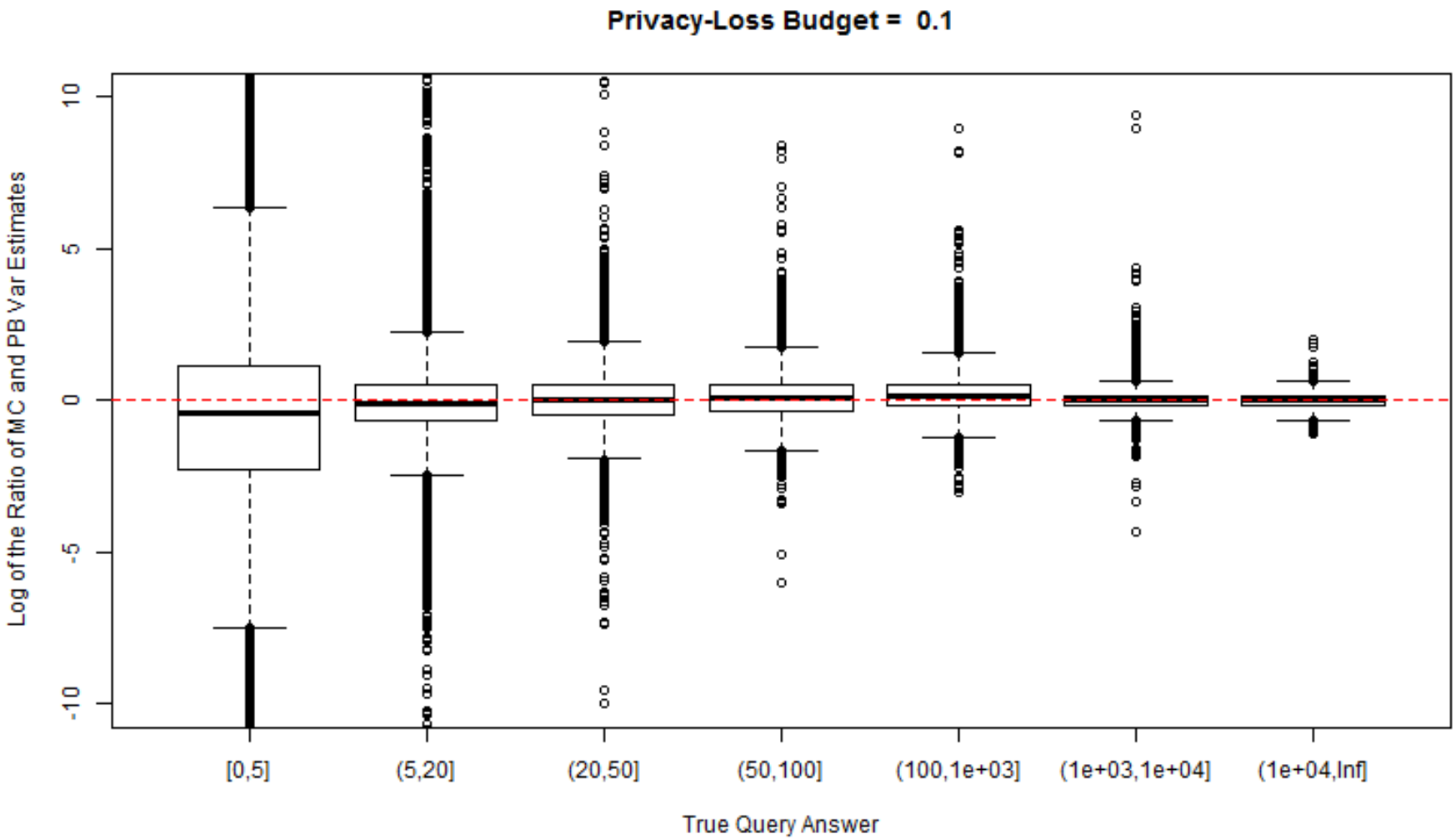
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results II



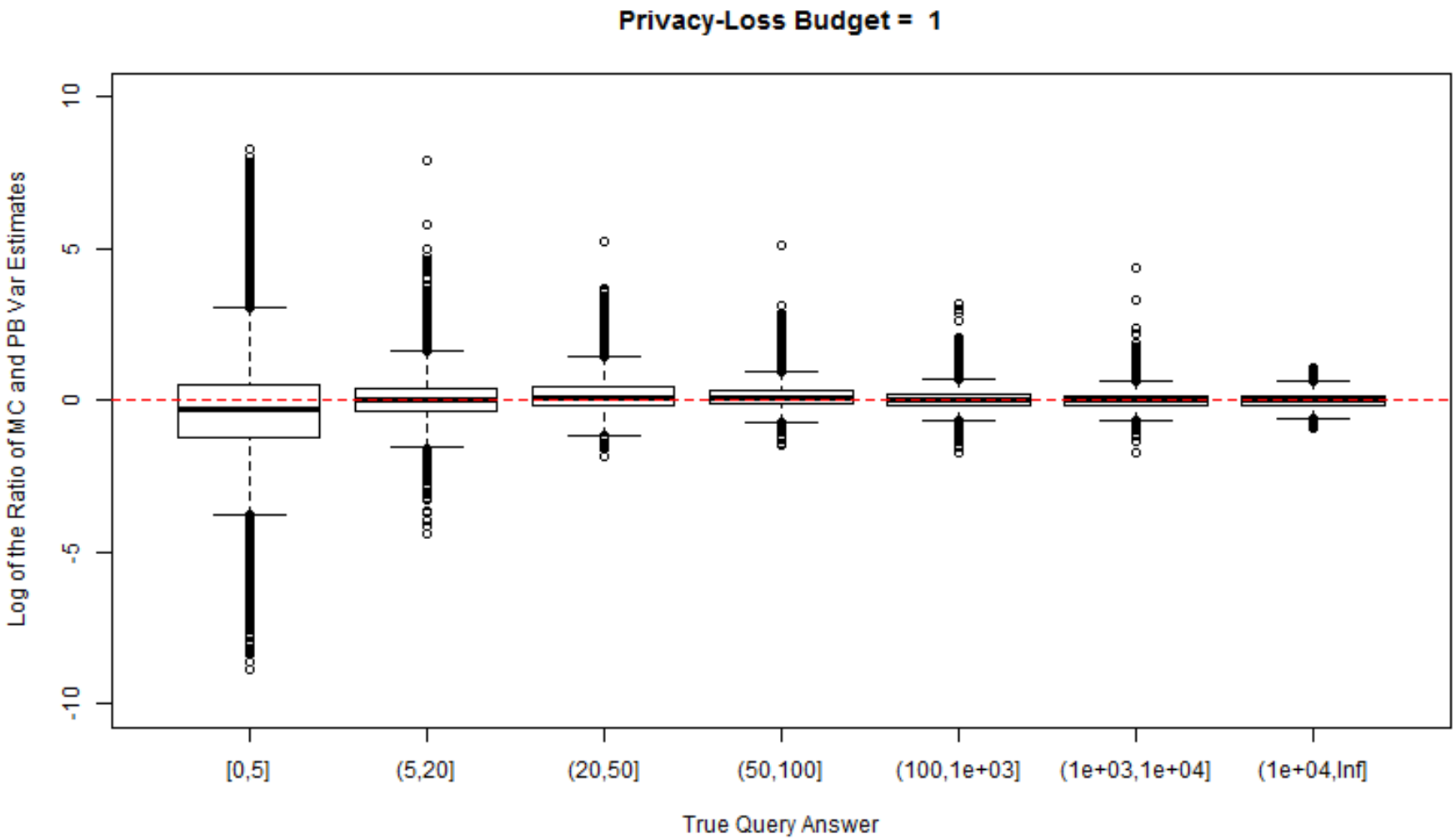
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results III



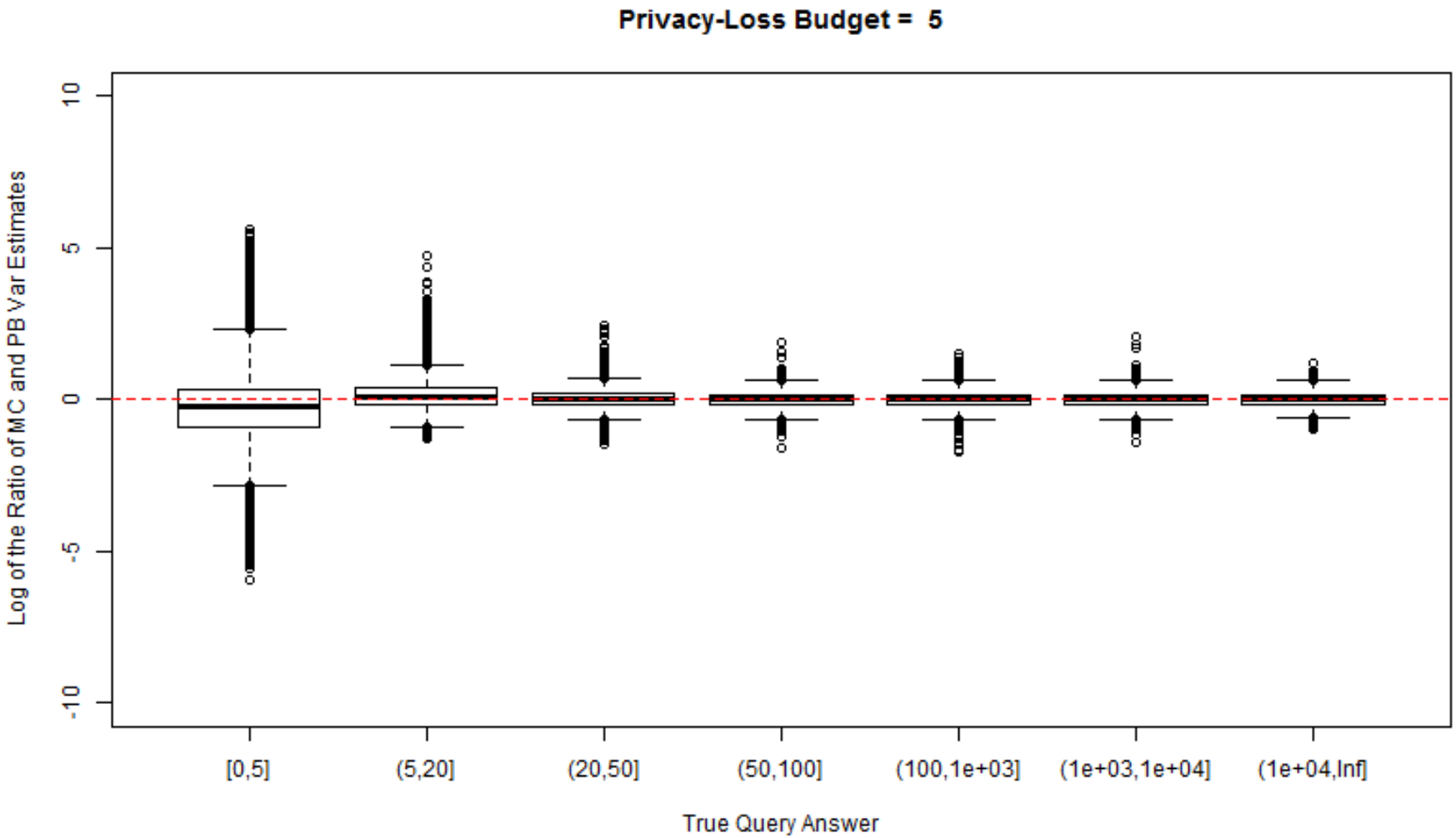
OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results III



OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Results III



OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER

Discussion

- The PB approach estimates the variance exceptionally well considering that it does not spend additional privacy-loss budget
- Does better for larger queries than smaller ones
- Improves with a larger privacy-loss budget
- In general, its success will be dependent on how well the initial DP estimate matches the truth
- Additional work is needed on sufficient number of runs

References

Li, C., Miklau, G., Hay, M., McGregor, A., Rastogi, V. (2015). The matrix mechanism: optimizing linear counting queries under differential privacy. The VLDB journal, 24(6), 757-781.

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. IPUMS USA: Version 8.0 Extract of 1940 Census for U.S. Census Bureau Disclosure Avoidance Research [dataset]. Minneapolis, MN: IPUMS, 2018. https://doi.org/10.18128/D010.V8.0.EXT1940USCB

Thanks!

robert.ashmead@osumc.edu

OHIO COLLEGES OF MEDICINE
GOVERNMENT RESOURCE CENTER