Supplemental Materials to "Statistical Properties of Sanitized Results from Differentially Private Laplace Mechanism with Univariate Bounding Constraints"

Fang Liu*

There are two ways defining two data sets differing by one record $\Delta(\mathbf{x}, \mathbf{x}') = 1$. In the first definition, referred to as "Def 1" below, the two data sets have the same sample size n, but one and only one record differs in at least one attributes; a substitution would make the two data sets identical. In the second definition, referred to as "Def 2" below, one data set has one more record that the other, so the sample sizes differ by 1 (one is n and the other $n-1$), and a deletion (or an insertion) would make the two data sets identical. We calculated the l_1 GS for some common statistics below in both ways, and the results turned out to be the same, as shown below. Without loss of generality (WLOS), we assume it is the first observation that differs in data sets $\mathbf{x}$ and $\mathbf{x}'$ in Def 1, and $\mathbf{x}'$ does not have the first row (x_1) compared to $\mathbf{x}$ in Def 2 in the following calculation. Def 2 is more intuitive from the perspective of interpreting global sensitivity GS and DP, but the calculation of GS under Def 1 in general is much simpler (when $\mathbf{x}$ and $\mathbf{x}'$ are of the same size) analytically than that under Def 2. In most cases, the two definitions lead to the same GS (such as mean, variance, and covariance); In the other two statistics (pooled variance and pooled covariance across multiple groups), the GS calculated under the two definitions are different, but the discrepancy between the two in terms of its impact on the sanitized results usually diminishes when n gets large.

Sections 1 to 5 presents the global sensitivity for mean, variance, and covariance; and Section 6 provides additional simulation results in the second simulation study on releasing a vector of proportions (a histogram).

1 l_1-GS of sample mean and proportion

Denote by $\delta_{\bar{x}}$ the GS of a sample mean of variable x that is globally bounded in $[c_0, c_1]$, then

$$\textbf{Def 1: } \delta_{\bar{x}} = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} \left| \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} x_i' \right| = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} n^{-1} \left| x_1 - x_1' \right| = n^{-1}(c_1 - c_0). \blacksquare$$

$$\textbf{Def 2: } \delta_{\bar{x}} = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} \left| \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i - \tfrac{1}{n-1}\sum_{i=2}^{n} x_i \right|$$

$$= \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} \left| n^{-1}((n-1)\bar{x}_- + x_n) - \bar{x}_- \right| = n^{-1} \sup_{\mathbf{x},\mathbf{x}':\Delta(x,x')=1} \left| x_n - \bar{x}_- \right|,$$

where $\bar{x}_- = (n-1)^{-1}\sum_{i=2}^{n} x_i$. The maximum possible value of $|x_n - \bar{x}|$ is $c_1 - c_0$ across all possible data sets $\mathbf{x}$ and all possible ways leading to $\Delta(\mathbf{x}, \mathbf{x}') = 1$.

Therefore, $\delta_{\bar{x}} = n^{-1}(c_1 - c_0)$. $\blacksquare$

*Fang Liu is Associate Professor in the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556 (‡E-mail: fang.liu.131@nd.edu). The work is supported by the NSF Grants 1546373, 1717417, and the University of Notre Dame Faculty Research Initiation Grant.

A sample proportion can be viewed as a special case of a sample mean with the mean operated on the indicator function with $c_1 = 1$ and $c_0 = 0$. Therefore, δ_1 of a single proportion is n^{-1}. In addition to releasing a single proportion, in many practical cases, it is of interest to release a whole histogram $\mathbf{H}$ or a whole vector of proportions $\mathbf{p}$, the GS of which differ between the two definitions of $\Delta(\mathbf{x}, \mathbf{x}')$. Specifically, in Def 1, where n is the same between $\mathbf{x}$ and $\mathbf{x}'$, $\delta_1 = 2$ and $\delta_1 = 2n^{-1}$ for $\mathbf{H}$ and $\mathbf{p}$, respectively; in Def 2, $\delta_1 = 1$ and $\delta_1 = n^{-1}$.

2 l_1-GS of sample variance

Denote the sample variances of $\mathbf{x}$ and $\mathbf{x}'$ by s^2 and s'^2, respectively, and the global bounds by $[c_0, c_1]$. The l_1 GS $\delta_{s^2} = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} |s^2 - s'^2|$, where

$$
\begin{aligned}
\textbf{Def 1: } s^2 - s'^2 &= (n-1)^{-1}\left(\textstyle\sum_{i=1}^n (x_i^2 - x_i'^2) - n\left(\bar{x}^2 - \bar{x}'^2\right)\right) \\
&= (n-1)^{-1}\left(x_1^2 - x_1'^2 - n^{-1}(x_1 - x_1')(2\textstyle\sum_{i=2}^n x_i + x_1 + x_1')\right) \\
&= (n-1)^{-1}(x_1 - x_1')\left(x_1 + x_1' - n^{-1}(2\textstyle\sum_{i=2}^n x_i + x_1 + x_1')\right) \\
&= (n-1)^{-1}(x_1 - x_1')\left((1-n^{-1})x_1 + (1-n^{-1})x_1' - 2n^{-1}\textstyle\sum_{i=2}^n x_i\right) \\
&= n^{-1}(x_1^2 - x_1'^2) - 2n^{-1}(x_1 - x_1')\bar{x}_-, \text{ where } \bar{x}_- = (n-1)^{-1}\textstyle\sum_{i=2}^n x_i \\
&= n^{-1}(x_1^2 - 2x_1\bar{x}_- + \bar{x}_-^2 - x_1'^2 + 2x_1'\bar{x}_- - \bar{x}_-^2) \\
&= n^{-1}(x_1 - \bar{x}_-)^2 - n^{-1}(x_1' - \bar{x}_-)^2. \qquad (1)
\end{aligned}
$$

where $\bar{x}_- = (n-1)^{-1}\sum_{i=2}^n x_i$. Since both terms in Eq (1) are ≥ 0, $|s^2 - s'^2|$ is maximized when $(x_1 - \bar{x}_-)^2$ reaches its maximum $(c_1 - c_0)^2$ and $(x_1' - \bar{x}_-)^2 = 0$, or when $(x_1 - \bar{x}_-)^2 = 0$ and $(x_1' - \bar{x}_-)^2$ reaches its maximum $(c_1 - c_0)^2$. Therefore,

$$
\delta_{s^2} = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} |s^2 - s'^2| = n^{-1}(c_1 - c_0)^2. \quad \blacksquare
$$

$$
\begin{aligned}
\textbf{Def 2: } s^2 - s'^2 &= \frac{x_1^2 + \sum_{i=2}^n x_i^2 - n^{-1}((n-1)\bar{x}_- + x_1)^2}{n-1} - \frac{\sum_{i=2}^n x_i^2 - (n-1)\bar{x}_-^2}{n-2} \\
&= [(n-1)(n-2)]^{-1}\left[(n-2)x_1^2 + (n-2)\textstyle\sum_{i=2}^n x_i^2\right. \\
&\left. -(n-2)n^{-1}\left((n-1)^2\bar{x}_-^2 + 2(n-1)x_1\bar{x}_- + x_1^2\right) - (n-1)\textstyle\sum_{i=2}^n x_i^2 + (n-1)^2\bar{x}_-^2\right] \\
&= \frac{-\sum_{i=2}^n x_i^2 + (n-2)(1-n^{-1})x_1^2 + 2n^{-1}(n-1)^2\bar{x}_-^2 - 2n^{-1}(n-1)(n-2)\bar{x}_-x_1}{(n-1)(n-2)} \\
&= -\frac{\sum_{i=2}^n x_i^2}{(n-1)(n-2)} + \frac{x_1^2}{n} + 2\frac{(n-1)}{n(n-2)}\bar{x}_-^2 - 2\frac{\bar{x}_-x_1}{n} \\
&= \frac{1}{n}\left[x_1^2 - 2\bar{x}_-x_1 + \bar{x}_-^2\right] - \frac{\bar{x}_-^2}{n} + \frac{2(n-1)\bar{x}_-^2}{n(n-2)} - \frac{\sum_{i=2}^n x_i^2}{(n-1)(n-2)} \\
&= \frac{1}{n}(x_1 - \bar{x}_-)^2 + \frac{\bar{x}_-^2}{n-2} - \frac{\sum_{i=2}^n x_i^2}{(n-1)(n-2)} \\
&= \frac{1}{n}(x_1 - \bar{x}_-)^2 - \frac{\sum_{i=2}^n x_i^2 - (n-1)\bar{x}_-^2}{(n-1)(n-2)} = n^{-1}(x_1 - \bar{x}_-)^2 - (n-1)^{-1}s'^2.
\end{aligned}
$$

$|s^2 - s'^2|$ is maximized in either of the following two cases, whichever is larger: 1) when $(x_1 - \bar{x}_-)^2 = 0$ and s'^2 reaches maximum, and 2) when $s'^2 = 0$ and $(x_1 - \bar{x}_-)^2$ reaches maximum. Case 1) occurs when

$x_1 = \bar{x}_-$ and $s'^2 = (c_1 - c_0)^2 \frac{n-1}{4(n-2)}$, the maximum possible value of the sample variance with sample size $(n-1)$ (Shiffler and Harsha, 1980), leading to $\sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} |s^2 - s'^2| = (c_1 - c_0)^2/(4n-8)$. Case 2) is realized when $(x_1 - \bar{x}_-)^2 = (c_1 - c_0)^2$, and $s'^2 = 0$ (when all $x_i = c_0$ for $i = 2, \ldots, n$ if $x_1 = c_1$, or when all $x_i = c_1$ for $i = 2, \ldots, n$ if $x_1 = c_0$), leading to $\sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1x} |s^2 - s'^2| = (c_1 - c_0)^2/n$. Since Case 2 is larger (for $n > 2$), then

$$\delta_{s^2} = n^{-1}(c_1 - c_0)^2. \quad \blacksquare$$

3 l_1-GS of sample covariance

Denote the sample covariance between x_1 and x_2 in $\mathbf{x}$ by s_{12} and that in $\mathbf{x}'$ by s'_{12} respectively. Denote the global bounds of x_1 and x_2 by $[c_{10}, c_{11}]$ and $[c_{20}, c_{21}]$, respectively.

$$\begin{aligned}
\textbf{Def 1}: \; s_{12}^2 - s_{12}'^2 &= (n-1)^{-1}[\textstyle\sum_{i=1}^n x_{i1}x_{i2} - n\bar{x}_1\bar{x}_2 - \sum_{i=1}^n x'_{i1}x'_{i2} + n\bar{x}'_1\bar{x}'_2] \\
&= (n-1)^{-1}[x_{11}x_{12} - x'_{11}x'_{21} - n^{-1}(x_{11} + (n-1)\bar{x}_{1-})(x_{12} + (n-1)\bar{x}_{2-}) \\
&\qquad + n^{-1}(x'_{11} + (n-1)\bar{x}_{1-})(x'_{21} + (n-1)\bar{x}_{2-})] \\
&= n^{-1}[x_{11}x_{12} - x'_{11}x'_{21} - x_{11}\bar{x}_{2-} - \bar{x}_{1-}x_{12} + x'_{11}\bar{x}_{2-} + \bar{x}_{1-}x'_{21}] \\
&= n^{-1}[x_{11}\underbrace{(x_{12} - \bar{x}_{2-})}_{\text{term 1}} + \bar{x}_{1-}\underbrace{(x'_{21} - x_{12})}_{\text{term 2}} + x'_{11}\underbrace{(\bar{x}_{2-} - x'_{21})}_{\text{term 3}}], \qquad (2)
\end{aligned}$$

where $\bar{x}_{k-} = (n-1)^{-1}\sum_{i=2}^n x_{ki}$ $(k = 1, 2)$. WLOS, assume that $x_{2-} \le x_{12} \le x'_{12}$, then term 1 and term $2 > 0$, term $3 < 0$. Eq (2) is maximized when $x_{11} = \bar{x}_{1-} = c_{11}$ and $\bar{x}_{1-} = c_{10}$, and Eq (2) is then written as $n^{-1}[c_{11}(x_{12} - \bar{x}_{2-}) + c_{11}(x'_{21} - x_{12}) + c_{10}(\bar{x}_{2-} - x'_{21})] = (c_{11} - c_{10})(x'_{21} - \bar{x}_{2-})$, which reaches its maximum when $x'_{21} - \bar{x}_{2-} = c_{21} - c_{20}$. Therefore,

$$\delta_{s_{12}} = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} |s_{12} - s'_{12}| = n^{-1}(c_{11} - c_{10})(c_{21} - c_{20}). \quad \blacksquare$$

Def 2: Denote the sample mean of $\mathbf{x}'$ with one less observation by $\bar{\mathbf{x}}$. The GS of the covariance is $\delta_{s_{12}} = \sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} |s_{12} - s'_{12}|$, where

$$\begin{aligned}
\Delta s_{12} - s'_{12} &= \frac{\sum_{i=2}^n x_{i1}x_{i2} + x_{11}x_{12} - n^{-1}[((n-1)\bar{x}_{1-} + x_{11})((n-1)\bar{x}_{2-} + x_{12})]}{n-1} \\
&\quad - \frac{\sum_{i=2}^n x_{i1}x_{i2} - (n-1)\bar{x}_{1-}\bar{x}_{2-}}{n-2} \\
&= \frac{1}{(n-1)(n-2)}[(n-2)\textstyle\sum_{i=2}^n x_{i1}x_{i2} + (n-2)x_{11}x_{12} - (n-1)\sum_{i=2}^n x_{i1}x_{i2} + (n-1)^2\bar{x}_{1-}\bar{x}_{2-} \\
&\qquad - (n-2)n^{-1}\left((n-1)^2\bar{x}_{1-}\bar{x}_{2-} + (n-1)\bar{x}_{1-}x_{12} + (n-1)x_{11}\bar{x}_{2-} + x_{11}x_{12}\right)] \\
&= \frac{1}{(n-1)(n-2)}[-\textstyle\sum_{i=2}^n x_{i1}x_{i2} + n^{-1}(n-1)(n-2)x_{11}x_{12} + 2n^{-1}(n-1)^2\bar{x}_{1-}\bar{x}_{2-} \\
&\qquad - n^{-1}(n-1)(n-2)(\bar{x}_{1-}x_{12} + x_{11}\bar{x}_{2-})] \\
&= -\frac{\sum_{i=2}^n x_{i1}x_{i2}}{(n-1)(n-2)} + \frac{x_{11}x_{12}}{n} + 2\frac{(n-1)}{n(n-2)}\bar{x}_{1-}\bar{x}_{2-} - \frac{\bar{x}_{1-}x_{12} + x_{11}\bar{x}_{2-}}{n} \\
&= \frac{1}{n}[x_{11}x_{12} - \bar{x}_{1-}x_{12} - x_{11}\bar{x}_{2-} + \bar{x}_{1-}\bar{x}_{2-}] - \frac{\bar{x}_{1-}\bar{x}_{2-}}{n} - \frac{\sum_{i=2}^n x_{i1}x_{i2}}{(n-1)(n-2)} + 2\frac{(n-1)}{n(n-2)}\bar{x}_{1-}\bar{x}_{2-} \\
&= \frac{1}{n}(x_{11} - \bar{x}_{1-})(x_{12} - \bar{x}_{2-}) + \frac{\bar{x}_{1-}\bar{x}_{2-}}{n-2} - \frac{\sum_{i=2}^n x_{i1}x_{i2}}{(n-1)(n-2)}
\end{aligned}$$

$$= n^{-1}(x_{11} - \bar{x}_{1-})(x_{12} - \bar{x}_{2-}) - (n-1)^{-1} s'_{12} \tag{3}$$

Both terms $(x_{11} - \bar{x}_{1-})(x_{12} - \bar{x}_{2-})$ and s'_{12} in Eq (3) can be > 0 or < 0 and depend on $\bar{x}_{1-}$ and $\bar{x}_{2-}$, making the determination of the maximum of $|s_{12} - s'_{12}|$ complicated. Rewrite Eq (3) as

$$\begin{aligned}\Delta =& n^{-1}\left(x_{11} - (n-1)^{-1}\textstyle\sum_{i=2}^{n} x_{i1}\right)\left(x_{12} - (n-1)^{-1}\textstyle\sum_{i=2}^{n} x_{i2}\right) + \\ & (n-2)^{-1}(n-1)^{-2}\textstyle\sum_{i=2}^{n} x_{i1}\sum_{i=2}^{n} x_{i2} - (n-2)^{-1}(n-1)^{-1}\textstyle\sum_{i=2}^{n} x_{i1}x_{i2},\end{aligned}$$

which suggests Δ is a linear function of x_{i1} and x_{i2} for all $i = 1, \ldots, n$, indicating that the maximum of $|s_{12} - s'_{12}|$ occurs at the corners of the vector $(x_{11}, \ldots, x_{n1}, x_{12}, \ldots, x_{n2})$. To simplify the algebra, we work with transformed x_{i1} and x_{i2}, that is, $x_{i1} = (c_{11} - c_{10})^{-1}(x_{i1} - c_{10})$ and $x_{i2} = (c_{21} - c_{20})^{-1}(x_{i2} - c_{20})$. After the transformation, x_{i1} and x_{i2} are both bounded within $[0, 1]$. The goal is to determine between 0 and 1 at each of the $2n$ position that lead to the maximum $|s_{12} - s'_{12}|$. The maximum Δ on the original scale can be obtained by scaling the maximum Δ on the $[0, 1] \times [0, 1]$ scale with $(c_{11} - c_{10})(c_{21} - c_{20})$.

Let $k_1 = \#\{x_{i1} = 1\}, k_2 = \#\{x_{i2} = 1\}$ and $k_3 = \sum_i x_{i1}x_{i2}$ for $i = 2, \ldots, n$. Thus $\bar{x}_1 = (n-1)^{-1}k_1$, and $\bar{x}_2 = (n-1)^{-1}k_2$. It is easy to see $k_3 \in [\max(0, k_1 + k_2 - (n-1)), \min(k_1, k_2)]$. WLOS, assume $k_2 \leq k_1$, thus $0 \leq k_1 \leq n-1, 0 \leq k_2 \leq k_1, \max(0, k_1 + k_2 - (n-1)) \leq k_3 \leq k_2$. Among the three, k_1 vary from 0 to $n-1$, while the range of k_2 depends on k_1, and that of k_3 depends on k_1 and k_2.

1. If $(x_{11}, x_{12}) = (0, 0)$, then

$$s_{12} - s'_{12} = \frac{k_1 k_2}{n(n-1)^2} - \frac{k_3 - (n-1)^{-1}k_1 k_2}{(n-1)(n-2)} = \frac{2k_1 k_2}{n(n-1)(n-2)} - \frac{k_3}{(n-1)(n-2)}$$

   $s_{12} - s'_{12}$ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

   - If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and $s_{12} - s'_{12} = 0$
   - If $k_1 = n-1$, then $k_2 \in [0, n-1], k_3 = k_2$, and $s_{12} - s'_{12} = \frac{k_2}{n(n-1)}$, thus the maximum of $|s_{12} - s'_{12}|$ is n^{-1} when $k_1 = k_2 = k_3 = n-1$

2. If $(x_{11}, x_{12}) = (0, 1)$, then

$$\begin{aligned}s_{12} - s'_{12} &= -\frac{k_1(n-1-k_2)}{n(n-1)^2} - \frac{k_3 - (n-1)^{-1}k_1 k_2}{(n-1)(n-2)} \\ &= \frac{2k_1 k_2}{n(n-1)(n-2)} - \frac{k_3}{(n-1)(n-2)} - \frac{k_1}{n(n-1)}\end{aligned}$$

   $s_{12} - s'_{12}$ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

   - If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and $s_{12} - s'_{12} = 0$
   - If $k_1 = n-1$, then $k_2 \in [0, n-1], k_3 = k_2$, and $s_{12} - s'_{12} = \frac{k_2}{n(n-1)} - \frac{1}{n}$. when $k_2 = 0$, $s_{12} - s'_{12} = -n^{-1}$, when $k_2 = (n-1)$, $s_{12} - s'_{12} = 0$, thus the maximum of $|s_{12} - s'_{12}|$ is n^{-1} when $k_1 = n-1, k_2 = k_3 = 0$

3. If $(x_{11}, x_{12}) = (1, 0)$, then

$$\begin{aligned}s_{12} - s'_{12} &= -\frac{k_2(n-1-k_1)}{n(n-1)^2} - \frac{k_3 - (n-1)^{-1}k_1 k_2}{(n-1)(n-2)} \\ &= \frac{2k_1 k_2}{n(n-1)(n-2)} - \frac{k_3}{(n-1)(n-2)} - \frac{k_2}{n(n-1)}\end{aligned}$$

   $s_{12} - s'_{12}$ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

- If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and $s_{12} - s'_{12} = 0$
- If $k_1 = n-1$, then $k_2 \in [0, n-1], k_3 = k_2$, and $s_{12} - s'_{12} = \frac{k_2}{n(n-1)} - \frac{k_2}{n(n-1)} = 0$.

4. If $(x_{11}, x_{12}) = (1, 1)$, then

$$\begin{aligned} s_{12} - s'_{12} &= -\frac{(n-1-k_1)(n-1-k_2)}{n(n-1)^2} - \frac{k_3 - (n-1)^{-1}k_1k_2}{(n-1)(n-2)} \\ &= \frac{1}{n} + \frac{2k_1k_2}{n(n-1)(n-2)} - \frac{k_3}{(n-1)(n-2)} - \frac{k_1}{n(n-1)} - \frac{k_2}{n(n-1)} \end{aligned}$$

$s_{12} - s'_{12}$ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

- If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and $s_{12} - s'_{12} = n^{-1}$, thusthe maximum of $|s_{12} - s'_{12}|$ is n^{-1} when $k_1 = k_2 = k_3 = 0$
- If $k_1 = n-1$, then $k_2 \in [0, n-1], k_3 = k_2$, and $s_{12} - s'_{12} = 0$.

Note that the above results are obtained under the assumption $k_2 \leq k_1$. The same sets of results are obtained by changing the assumption to $k_1 \leq k_1$, and flipping the labels the two variables in each of the 4 cases above. In summary, on the transformed scale $[0,1] \times [0,1]$, the maximum of $|s_{12} - s'_{12}|$ is n^{-1}, that occurs if any one of the following conditions holds: 1) $x_{11} = x_{12} = 0$, and $x_{i1} = 1, x_{i2} = 1$ for all $i = 2, \ldots, n$; 2) $x_{11} = x_{12} = 1$, and $x_{i1} = 0, x_{i2} = 0$ for all $i = 2, \ldots, n$; 3) $x_{11} = 0, x_{12} = 1$, and $x_{i1} = 1$ and $x_{i2} = 0$ for all $i = 2, \ldots, n$; 4) $x_{11} = 1, x_{12} = 0$, and $x_{i1} = 0$ and $x_{i2} = 1$ for all $i = 2, \ldots, n$; Transforming back to the original scale $[c_{10}, c_{11}] \times [c_{20}, c_{21}]$, we have

$$\sup_{\mathbf{x},\mathbf{x}':\Delta(\mathbf{x},\mathbf{x}')=1} |s_{12} - s'_{12}| = n^{-1}(c_{11} - c_{10})(c_{21} - c_{20}),$$

which occurs if any one of the following conditions holds: 1) $x_{11} = c_{10}$, $x_{12} = c_{20}$, and $x_{i1} = c_{11}, x_{i2} = c_{21}$ for all $i = 2, \ldots, n$; 2) $x_{i1} = c_{11}, x_{i2} = c_{21}$, and $x_{i1} = c_{10}, x_{i2} = c_{20}$ for all $i = 2, \ldots, n.$; 3) $x_{11} = c_{10}, x_{12} = c_{21}$, and $x_{i1} = c_{11}$ and $x_{i2} = c_{20}$ for all $i = 2, \ldots, n$; and 4) $x_{11} = c_{11}, x_{12} = c_{20}$, and $x_{i1} = c_{10}$ and $x_{i2} = c_{21}$ for all $i = 2, \ldots, n$;

4 l_1-GS pooled sample variance

Denote the number of cells of J and n_j is the number of observations in cell j (for $j = 1, \ldots, J$). Assume each cell contributes to the pooled variance s_p^2; in other words, there are at least 2 observations in each cell ($n_j \geq 2$ for $j = 1, \ldots, J$). Denote the total sample size by $n = \sum_{j=1}^{J} n_j$, then

$$s_p^2 = (n-J)^{-1} \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \right),$$

where x_{ij} is the i-th observation in cell j, and $\bar{x}_j$ is the mean of cell j.

Def 1: WLOS, suppose it is the first observation in cell $j = 1$ that differs between data $\mathbf{x}$ and $\mathbf{x}'$, then

$$\begin{aligned} \Delta = s_p^2 - s_p'^2 &= (n-J)^{-1} \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \right) - (n-J)^{-1} \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} (x'_{ij} - \bar{x'}_j)^2 \right) \\ &= (n-J)^{-1}(n_1 - 1) \underbrace{(n_1-1)^{-1} \left(\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 - \sum_{i=1}^{n_1} (x_{i1} - \bar{x}'_1)^2 \right)}_{\text{term 1}}. \end{aligned}$$

Term 1 is the difference in the variance in cell 1 between $\mathbf{x}$ and $\mathbf{x}$, and its maximum is $n_1^{-1}(c_1 - c_0)^2$ per the results in Section 2. Therefore, $\max|\Delta| = (n-J)^{-1}(c_1 - c_0)^2(1 - n_1^{-1})$, which reaches maximum

if n_1 is the largest among $(n_1, \ldots, n_J)$. All taken together, the GS of s_p^2 is

$$\delta_{s_p^2} = (c_1 - c_0)^2 (n-J)^{-1} \left(1 - n_{\max}^{-1}\right), \text{ which can be approximated by}$$
$$\delta_{s_p^2} = (c_1 - c_0)^2 (n-J)^{-1}$$

if $n_{\max}$ is large or when $n_{\max}$ itself cannot be released without sanitization. ■

Def 2: WLOS, suppose the first observation in cell 1 is removed in $\mathbf{x}'$ compared to $\mathbf{x}$, then

$$s_{p-}^2 = (n-1-J)^{-1} \left(\sum_{j=2}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{i=2}^{n_1} (x_{i1} - \bar{x}_{1-})^2 \right),$$

where $\bar{x}_{1-}$ is the mean of cell 1 without the first observation. Let

$$\begin{aligned}
\Delta = s_p^2 - s_{p-}^2 &= \frac{(n-1-J)\left(\sum_{j=2}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 \right)}{(n-J)(n-1-J)} \\
&\quad - \frac{(n-J)\left(\sum_{j=2}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 + \sum_{i=2}^{n_1} (x_{i1} - \bar{x}_{1-})^2 \right)}{(n-J)(n-1-J)} \\
&= \frac{(n-1-J)\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 - (n-J)\sum_{i=2}^{n_1}(x_{i1} - \bar{x}_{1-})^2 - \sum_{j=2}^{J}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{(n-J)(n-1-J)} \\
&= \frac{(n-1-J)\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 - (n-J)\sum_{i=2}^{n_1}(x_{i1} - \bar{x}_{1-})^2 + \sum_{i=2}^{n_1}(x_{i1} - \bar{x}_{1-})^2}{(n-J)(n-1-J)} \\
&\quad - \frac{\sum_{i=2}^{n_1}(x_{i1} - \bar{x}_{1-})^2 + \sum_{j=2}^{J}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2}{(n-J)(n-1-J)} \\
&= \frac{\left(\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 - \sum_{i=2}^{n_1}(x_{i1} - \bar{x}_{1-})^2\right)}{(n-J)} - \frac{s_{p-}^2}{n-J} \\
&= \frac{\sum_{i=1}^{n_1} x_{i1}^2 - \frac{((n_1-1)\bar{x}_{1-} + x_{11})^2}{n_1} - \sum_{i=2}^{n_1} x_{i1}^2 + (n_1-1)\bar{x}_{1-}^2 - s_{p-}^2}{n-J} \\
&= \frac{(1 - n_1^{-1})x_{11}^2 + (1 - n_1^{-1})\left(\bar{x}_{1-}^2 - 2\bar{x}_1 x_{11}\right) - s_{p-}^2}{n-J} = \frac{(1 - n_1^{-1})(x_{11} - \bar{x}_{1-})^2 - s_{p-}^2}{n-J}
\end{aligned}$$

Since $(x_{11} - \bar{x}_{1-})^2 \geq 0$ and $s_{p-}^2 \geq 0$, the maximum of $|\Delta|$ takes the larger value of the two: $\max\{(x_{11} - \bar{x}_1)^2\}$ (case 1; $s_{p-}^2 = 0$ in this case), and $\max\{s_{p-}^2\}$ (case 2; $(x_{11} - \bar{x}_{1-})^2 = 0$.

- case 1: $(x_{11} - \bar{x}_1)^2 = (c_1 - c_0)^2$ when $(x_{11}, \bar{x}_{1-}) = (c_0, c_1)$ or (c_1, c_0). If $s_{p-}^2 = 0$, then the observations in a cell $j > 1$ are the same, and all the observations in cell 1 (without the first case) are the same (if $\bar{x}_{1-} = c_0$, then $x_{i1} \equiv c_0$ for $i = 2, \ldots, n_1$; if $\bar{x}_{1-} = c_1$, then $x_{i1} \equiv c_1$ for $i = 2, \ldots, n_1$). Therefore, $\max|\Delta| = (1 - n_1^{-1})(c_1 - c_0)^2/(n-J)$, which is again is maximized when n_1 is the maximum among all cell sizes. That is, $\max|\Delta| = (1 - n_{\max}^{-1})(c_1 - c_0)^2/(n-J)$.

- case 2: s_{p-}^2 reaches its maximum if the sum of squares of x in each cell reaches its maximum, which is $(c_1 - c_0)^2 n_j/4$ in cell $j \geq 2$ and $(c_1 - c_0)^2(n_1 - 1)/4$ in cell 1. Therefore, $\max\{s_{p-}^2\} = \frac{(c_1-c_0)^2}{4}\frac{(n_1-1)+\sum_{j=2}^{J} n_j}{n-1-J} = \frac{(n-1)(c_1-c_0)^2}{4(n-1-J)}$ and $\max|\Delta| = (c_1 - c_0)^2(n-1)/(4(n-1-J)(n-J))$.

To compare $\max|S|$ between case 1 and case 2,

$$\begin{aligned}
(1 - n_{\max}^{-1})(c_1 - c_0)^2/(n-J) &\quad * \quad (c_1 - c_0)^2(n-1)/(4(n-1-J)(n-J)) \\
(1 - n_{\max}^{-1}) &\quad * \quad (n-1)/(4(n-1-J)
\end{aligned}$$

$$n_{\max} \quad * \quad 4(n-1-J)/(3(n-1)-4J) = 1+(3-\tfrac{4J}{n-1})^{-1} \tag{4}$$

The right hand side (RHS) in Eq (4) < the left hand side (LHS) $n_{\max}$ except when $n = 2J$ (exactly 2 observations per cell in $\mathbf{x}$), which rarely happens in real-life data. Therefore,

$$\delta_{s_p} = \begin{cases} (c_1-c_0)^2 \frac{1-n_{\max}^{-1}}{n-J} & \text{if } n > 2J \\ (c_1-c_0)^2 \frac{n-1}{n(n-2)} & \text{if } n = 2J \text{ (exactly 2 observations per cell)} \end{cases},$$

which is approximated by

$$\delta_{s_p} = \begin{cases} (c_1-c_0)^2 (n-J)^{-1} & \text{if } n > 2*J \\ (c_1-c_0)^2 \frac{n-1}{n(n-2)} & \text{if } n = 2*J \text{ (exactly 2 observations per cell)} \end{cases} \blacksquare$$

if $n_{\max}$ is large or when $n_{\max}$ itself cannot be released without sanitization.

5 l_1-GS of pooled sample covariance

Denote the number of cells by J and the number of observations in cell j by n_j ($j = 1, \ldots, J$). Assume each cell contributes to the pooled covariance cov_p; that is, each cells has at least 2 observations ($n_j \geq 2$ for $j = 1, \ldots, J$). Denote total sample size by $n = \sum_{j=1}^{J} n_j$, then the pooled covariance between variables x and y is

$$\text{cov}_p = \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{n - J}$$

Def 1: WLOS, suppose it is the 1st observation in cell $j = 1$ that differs between two data sets, then

$$\begin{aligned} \Delta &= \text{cov}_p - \text{cov}_p = (n-J)^{-1} \left(\sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j) - \sum_{j=1}^{J} \sum_{i=1}^{n_j} (x'_{ij} - \bar{x}'_j)(y'_{ij} - \bar{y}'_j) \right) \\ &= (n-J)^{-1}(n_1 - 1) \underbrace{(n_1-1)^{-1} \left(\sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)(y_{ij} - \bar{y}_j) - \sum_{i=1}^{n_1} (x_{i1} - \bar{x}'_1)(y'_{ij} - \bar{y}'_j) \right)}_{\text{term 1}}. \end{aligned}$$

Term 1 is the difference of the sample covariance in cell 1 between $\mathbf{x}$ and $\mathbf{x}'$, the maximum of which is $n_1^{-1}(c_{11} - c_{10})(c_{21} - c_{20})$ per the results in Section 3. Therefore, $\max|\Delta| \leq (n-J)^{-1} \frac{(n_1-1)}{n_1} (c_{11} - c_{10})(c_{21} - c_{20})^2 = (n-J)^{-1}(1 - n_1^{-1})(c_1 - c_0)^2$, which again reaches the maximum if the n_1 is the largest among $(n_1, \ldots, n_J)$. All taken together, the GS of s_p^2 is

$$\delta_{s_p^2} = (c_{11} - c_{10})(c_{21} - c_{20})(n-J)^{-1} \left(1 - n_{\max}^{-1}\right), \text{ which can be approximated by}$$
$$\delta_{s_p^2} = (c_{11} - c_{10})(c_{21} - c_{20})(n-J)^{-1}$$

if $n_{\max}$ is large or when $n_{\max}$ itself cannot be released without sanitization. $\blacksquare$

Def 2: WLOS, suppose it is the 1st observation in cell $j = 1$ that is removed, then

$$\text{cov}_{p-} = \frac{\sum_{j=2}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j) + \sum_{i=2}^{n_1} (x_{i1} - \bar{x}_{1-})(y_{i1} - \bar{y}_{1-})}{n - 1 - J}$$

$$
\begin{aligned}
\text{Let } \Delta = \text{cov}_p - \text{cov}_{p-} &= \frac{(n-1-J)\left(\sum_{j=2}^{J}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)(y_{ij}-\bar{y}_j)+\sum_{i=1}^{n_1}(x_{i1}-\bar{x}_1)(y_{i1}-\bar{y}_1)\right)}{(n-J)(n-1-J)} \\
&\quad - \frac{(n-J)\left(\sum_{j=2}^{J}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)(y_{ij}-\bar{y}_j)+\sum_{i=2}^{n_1}(x_{i1}-\bar{x}_{1-})(y_{i1}-\bar{y}_{1-})\right)}{(n-J)(n-1-J)} \\
&= \frac{\sum_{i=1}^{n_1}(x_{i1}-\bar{x}_1)(y_{i1}-\bar{y}_1) - \sum_{i=2}^{n_1}(x_{i1}-\bar{x}_{1-})(y_{i1}-\bar{y}_{1-}) - \text{cov}_{p-}}{(n-J)} \\
&= \frac{\sum_{i=1}^{n_1}x_{i1}y_{i1} - n_1^{-1}((n_1-1)\bar{x}_{1-}+x_{11})((n_1-1)\bar{y}_{1-}+y_{11}) - \sum_{i=2}^{n_1}x_{i1}y_{i1} + (n_1-1)\bar{x}_{1-}\bar{y}_{1-} - \text{cov}_{p-}}{(n-J)} \\
&= \frac{(1-n_1^{-1})x_{11}y_{11} + (1-n_1^{-1})\left(\bar{x}_{1-}\bar{y}_{1-} - \bar{x}_{1-}x_{11} - \bar{y}_{1-}y_{11}\right) - \text{cov}_{p-}}{(n-J)} \\
&= \frac{(1-n_1^{-1})(x_{11}-\bar{x}_{1-})(y_{11}-\bar{y}_{1-}) - \text{cov}_{p-}}{(n-J)} \\
&= \underbrace{\frac{(1-n_1^{-1})(x_{11}-\bar{x}_{1-})(y_{11}-\bar{y}_{1-}) - \frac{\sum_{i=2}^{n_1}(x_{i1}-\bar{x}_{1-})(y_{i1}-\bar{y}_{1-})}{(n-1-J)}}{n-J}}_{\text{term 1}} - \underbrace{\frac{\sum_{j=2}^{J}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)(y_{ij}-\bar{y}_j)}{(n-1-J)(n-J)}}_{\text{term 2}}
\end{aligned}
$$

Term 1 is independent from term 2, meaning how term 1 changes does not the affect of the value of term 2. $\max|\Delta|$ occurs when 1) term 1 reaches the maximum and term 2 is at its minimum, or 2) term 1 reaches its minimum and term 2 is at its maximum, whichever is larger gives $\max\Delta$. The maximum and minimum of term 1 can be obtained as follows. First, We transform x and y by $x = (c_{11}-c_{10})^{-1}(x-c_{10})$ and $y = (c_{21}-c_{20})^{-1}(y-c_{20})$. The transformed variables range from $[0,1]$. Let $k_1 = \#\{x_{i1}=1\}, k_2 = \#\{y_{i1}=1\}$ and $k_3 = \sum_i x_{1i}y_{i1}$ for $i = 2,\ldots,n_1$ in cell 1. Thus $\bar{x} = (n_1-1)^{-1}k_1$, and $\bar{y} = (n_1-1)^{-1}k_2$. It is easy to obtain that $k_3 \in [\max(0, k_1+k_2-(n_1-1)), \min(k_1,k_2)]$. WLOS, assume $k_2 \le k_1$, thus $0 \le k_1 \le n_1-1, 0 \le k_2 \le k_1, \max(0,k_1+k_2-(n_1-1)) \le k_3 \le k_2$. Therefore, the range of the values that k_2 depends on k_1, and that of k_3 depends on k_1 and k_2.

- If $(x_{11},y_{11}) = (0,0)$, then term 1 is

$$
\begin{aligned}
&\frac{(1-n_1^{-1})(n_1-1)^{-2}k_1k_2 - \frac{k_3-(n_1-1)^{-1}k_1k_2}{n-1-J}}{(n-J)} \\
=&\frac{(n-J+n_1-1)k_1k_2}{n_1(n_1-1)(n-J)(n-1-J)} - \frac{k_3}{(n-J)(n-1-J)}
\end{aligned}
$$

  Δ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1,k_2,k_3).

  - If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and term 1 is 0
  - If $k_1 = n_1-1$, then $k_2 \in [0, n_1-1], k_3 = k_2$, and the term 1 is $\frac{k_2}{n_1(n-J)}$, thus its maximum is $\frac{n_1-1}{n_1(n-J)}$ when $k_2 = n_1-1$; and its minimum is 0 when $k_2 = 0$

  Therefore, when $(x_{11},y_{11}) = (c_{10},c_{20})$, then term 1 is $\in (c_{11}-c_{10})(c_{21}-c_{20})\left[0, \frac{1-n_1^{-1}}{n-J}\right]$

- If $(x_{11},y_{11}) = (0,1)$, then term 1 is

$$
\frac{-(1-n_1^{-1})(n_1-1)^{-2}k_1(n_1-1-k_2) - \frac{k_3-(n_1-1)^{-1}k_1k_2}{n-1-J}}{(n-J)}
$$

$$=\frac{(n-J+n_1-1)k_1k_2}{n_1(n_1-1)(n-J)(n-1-J)}-\frac{k_3}{(n-J)(n-1-J)}-\frac{k_1}{n_1(n-J)}$$

Δ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

- If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and term 1 is 0
- If $k_1 = n_1 - 1$, then $k_2 \in [0, n_1 - 1], k_3 = k_2$, and the term 1 is $\frac{k_2}{n_1(n-J)} - \frac{n_1-1}{n_1(n-J)}$, thus its maximum is 0 when $k_2 = n_1 - 1$; and its minimum is $-\frac{n_1-1}{n_1(n-J)}$ when $k_2 = 0$

Therefore, when $(x_{11}, y_{11}) = (c_{11}, c_{20})$, then term 1 is $(c_{11} - c_{10})(c_{21} - c_{20}) \in \left[\frac{n_1^{-1}-1}{n-J}, 0\right]$

- If $(x_{11}, y_{11}) = (1, 0)$, then

$$\frac{-(1-n_1^{-1})(n_1-1)^{-2}k_2(n_1-1-k_1) - \frac{k_3-(n_1-1)^{-1}k_1k_2}{n-1-J}}{(n-J)}$$
$$=\frac{(n-J+n_1-1)k_1k_2}{n_1(n_1-1)(n-J)(n-1-J)}-\frac{k_3}{(n-J)(n-1-J)}-\frac{k_2}{n_1(n-J)}$$

Δ is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

- If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and term 1 is 0
- If $k_1 = n_1 - 1$, then $k_2 \in [0, n_1 - 1], k_3 = k_2$, and the term 1 also becomes 0 regardless of the value of k_2

Therefore, when $(x_{11}, y_{11}) = (c_{10}, c_{21})$, then term 1 is 0

- If $(x_{11}, y_{11}) = (1, 1)$, then

$$\frac{(1-n_1^{-1})(n_1-1)^{-2}(n_1-1-k_1)(n_1-1-k_2) - \frac{k_3-(n_1-1)^{-1}k_1k_2}{n-1-J}}{(n-J)}$$
$$=\frac{1-n_1^{-1}}{n-J}+\frac{(n-J+n_1-1)k_1k_2}{n_1(n_1-1)(n-J)(n-1-J)}-\frac{k_3}{(n-J)(n-1-J)}-\frac{k_1}{n_1(n-J)}-\frac{k_2}{n_1(n-J)}$$

S is linear in k_1, k_2 and k_3, thus the minimum/maximum occurs at corners of (k_1, k_2, k_3).

- If $k_1 = 0$, then $k_2 = 0, k_3 = 0$, and term 1 becomes $\frac{n_1-1}{n_1(n-J)}$
- If $k_1 = n_1 - 1$, then $k_2 \in [0, n_1 - 1], k_3 = k_2$, and term 1 becomes 0 regardless of k_2

Therefore, when $(x_{11}, y_{11}) = (c_{11}, c_{21})$, then term 1 is $\in (c_{11} - c_{10})(c_{21} - c_{20}) \left[0, \frac{1-n_1^{-1}}{n-J}\right]$

Taken together, term $1 \in \left[-\frac{1-n_1^{-1}}{n-J}, \frac{1-n_1^{-1}}{n-J}\right]$. For term 2, the numerator $\sum_{j=2}^{J}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j) = \sum_{j=2}^{J}(n_j - 1)\text{cov}_j$. Since $-\sum_{j=2}^{J}(n_j - 1)\sqrt{\text{var}(x_j)\text{var}(y_j)} \le \sum_{j=2}^{J}(n_j - 1)\text{cov}_j \le \sum_{j=2}^{J}(n_j - 1)\sqrt{\text{var}(x_j)\text{var}(y_j)}$, so $-(c_{11}-c_{10})(c_{21}-c_{20})\frac{n-n_1}{4} \le \sum_{j=2}^{J}(n_j-1)\text{cov}_j \le (c_{11}-c_{10})(c_{21}-c_{20})\frac{n-n_1}{4}$. Terms 1 and 2 taken together, $\Delta \in (c_{11}-c_{10})(c_{21}-c_{20}) \times \left[-\left(\frac{n_1-1}{n_1(n-J)} + \frac{n-n_1}{4(n-J)(n-n_1-J+1)}\right), \frac{n_1-1}{n_1(n-J)} + \frac{n-n_1}{4(n-J)(n-1-J)}\right]$, and the maximum of $|\Delta|$ is

$$(n-J)^{-1}(c_{11}-c_{10})(c_{21}-c_{20})\left(1-\frac{1}{n_1}+\frac{n-n_1}{4(n-1-J)}\right), \tag{5}$$

which reaches the maximum at $n_1 = 2\sqrt{n-1-J}$, being plugging back in Eq (5), we have

$$\delta_{\text{COV}_p} = \frac{(c_{11}-c_{10})(c_{21}-c_{20})}{n-J}\left(1+\frac{n}{4(n-1-J)}-\frac{1}{\sqrt{n-J-1}}\right)$$

Of course, n_1 being exactly $= 2\sqrt{n-1-J}$ is likely not to occur in real life since $2\sqrt{n-1-J}$ is most likely to be fractional, so δ_{COV_p} as given above is not an exact bound.

6 Additional Results from Simulation Study 2

Figures S1 to S2 on pages 11-12 of this document present the simulation results regarding the inferences of some linear combinations of $\mathbf{p}$ based on the synthetic data from the rescaling and the universal histogram approaches. A brief discussion on the results are presented in the main manuscript.

Reference

Shiffler, R. E. and Harsha, P. D. (1980), Upper and lower bounds for the sample standard deviation, Teaching Statistics, 2(3):84-86

Hay, M., Rastogiz, V., Miklauy, G. and Suciu, D. (2010) Boosting the Accuracy of Differentially Private Histograms Through Consistency, Proceedings of the VLDB Endowment, 3(1): 1021-1032

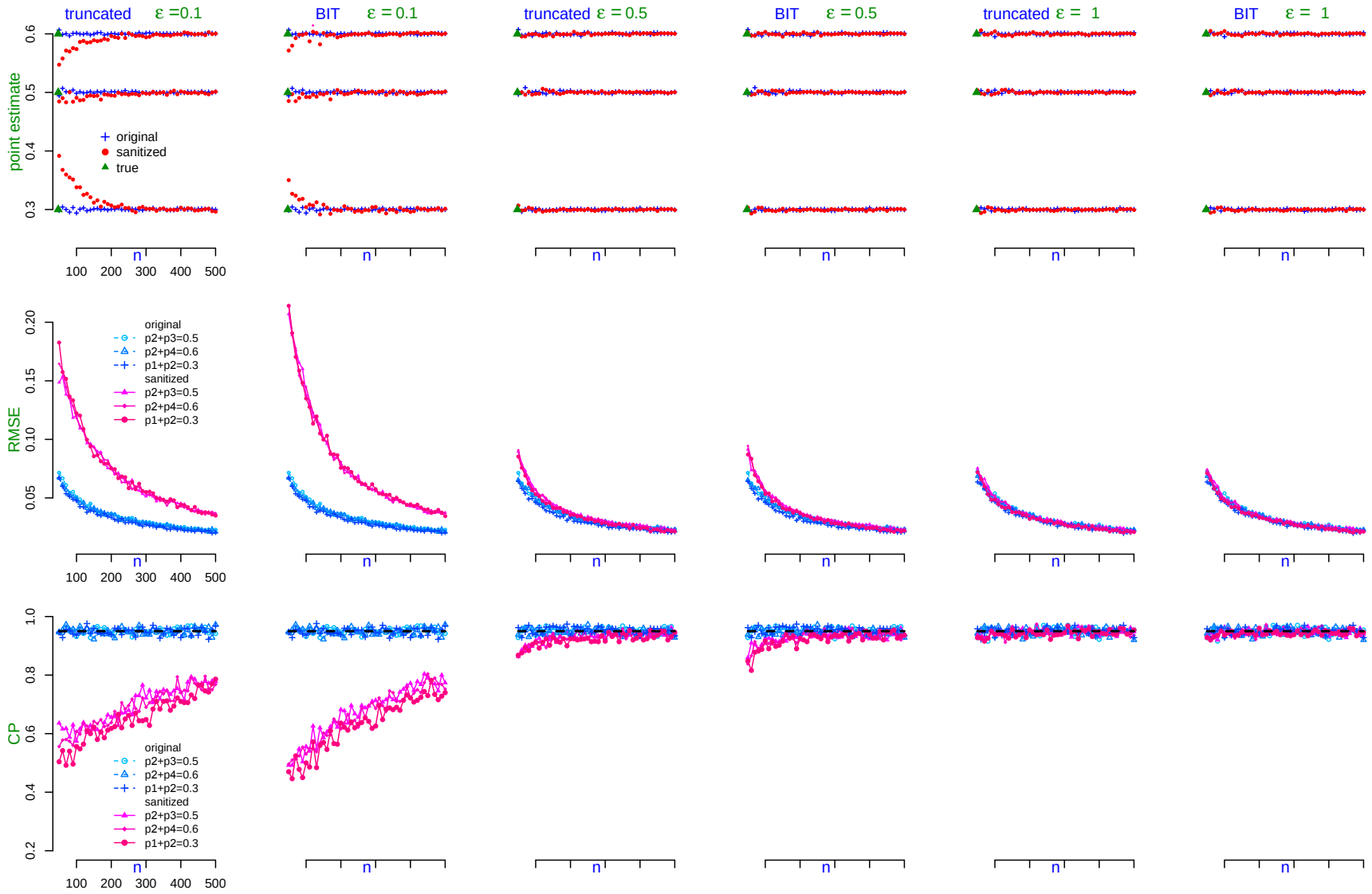

Figure S1: Bias, RMSE and CP of sanitized proportions in the rescaling approach (red lines represent the original linear combinations $(p_1+p_2, p_1+p_3, p_1+p_4)$ *of* $\mathbf{p}$*, and blue lines represent the sanitized versions)*

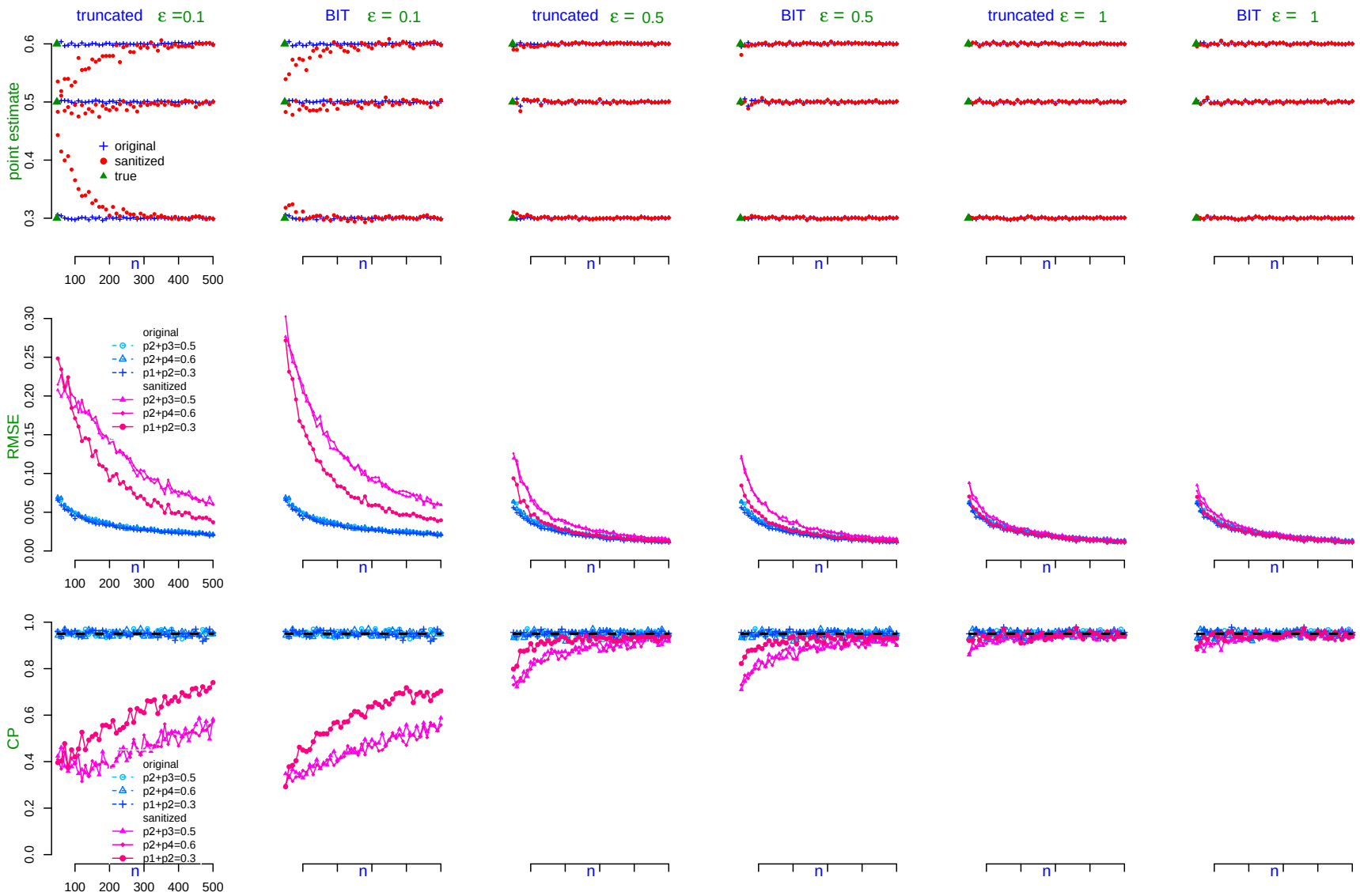

Figure S2: Bias, RMSE and CP of sanitized proportions in a modified universal histogram procedure based on Hay et. al. (2010) (red lines represent the original linear combinations $(p_1 + p_2, p_1 + p_3, p_1 + p_4)$ of $\mathbf{p}$, and blue lines represent the sanitized versions)