*International Journal of Statistical Sciences* *ISSN 1683-5603*
*Vol. 26(1), March, 2026, pp 65-82*
*DOI: https://doi.org/10.3329/ijss.v26i1. 88827*


# Inference about a Binomial Proportion under Privacy Protection

**Adam Hall[1], Nitul Singha[2] and Bimal Sinha[1,*]**

[1]Center for Statistical Research and Methodology, US Census Bureau, USA
Emails: adam.c.hall@census.gov; bimal.sinha@census.gov
[2]Department of Mathematics, Clarkson University, USA
Email: singhan@clarkson.edu
[*]Correspondence should be addressed to Bimal Sinha
(Email: bimal.sinha@census.gov)

[Received December 3, 2025; Accepted January 1, 2026]

## Abstract

In this paper we consider the inferential problem for a Binomial proportion in situations when the exact number of units possessing an attribute under consideration is unavailable due to privacy reasons; however a synthesized version of this number is available. The inference problem is addressed under three types of available information: noise added version and plug-in sampling based and posterior sampling based data. A comparison of the three modes of data source is made based on inferential accuracy and a measure of privacy.[1]

**Keywords:** Binomial proportion, Noise addition, Plug-in sampling, Posterior predictive sampling, Privacy, Synthetic data.

**AMS Classification:** 62F15, 62F30.

## 1. Introduction

Drawing valid statistical inference based on privacy protected data has been the topic of rigorous research at the US Census Bureau and other statistical agencies. The need to develop appropriate statistical methods based on perturbed or synthetic data rather than the original data stems from the fact that sometimes the original microdata are sensitive and cannot be released due to privacy considerations. What the statistical agencies will release instead is a synthetic version of the original data, hiding any confidential or sensitive parts, and also a valid method of data analysis based on such perturbations.

Pioneered by Rubin (1987, 1993, 1996) and subsequently developed by others, there is a substantial literature on statistical methods to hide the original sensitive data, and also to analyze a perturbed version of it (Drechsler and Reiter (2010); Drechsler (2011); Kinney et al. (2011); Kinney et al. (2014); Lin and Wise (2012); Little et al. (1993); Meng (1994); Klein et al. (2014); Klein and Sinha (2013a); Klein and Sinha (2015); Klein and Sinha (2016); Raghunathan et al. (2003); Reiter (2003); Reiter (2004); Reiter (2005a); Reiter (2005b); Reiter (2005c); Reiter and Kinney (2012); Reiter and

[1]Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau.

Mitra (2009); Reiter and Raghunathan (2007); Rubin (1987); Rubin (1993); Rubin (1996); Nayak et al. (2011); Sinha et al. (2011); Klein and Sinha (2013b)).

There are several ways to perturb an original dataset, most notably by using noise addition/multiplication, and two kinds of synthetic data: plug-in based (PLS) and posterior prediction based (PPS). Measures to study how successful the above methods are at protecting privacy have also been developed. Most applications are based on continuous attributes, often assuming normality.

In this paper we consider a situation when the response is discrete, and results in a Binomial sample $X \sim Binomial(n, \theta)$. Our goal is to draw valid inference about $\theta$ when $X$ is sensitive and cannot be directly released or used. There are many practical examples of this scenario (e.g., medical trial data, household survey results, etc.).

We consider three variations of $X$: (a) version with noise added, (b) synthetic version based on PLS of $X$, and (c) synthetic version based on PPS of $X$.

The paper is organized as follows. In Section 2, we provide a general description of the problem and briefly review the literature, pointing out some drawbacks of an existing method. Details based on noise addition are developed in Section 3, while Section 4 describes necessary inferential procedures based on PLS and PPS. Inference based on multiple imputations is discussed in Section 5. A standard measure of privacy protection is also evaluated for each method. Section 6 provides a comparison of the three methods of data perturbation and some concluding remarks.

## 2. Description of the research problem

Statistical agencies around the world regularly conduct surveys to collect pertinent information on many aspects of the nation such as topics related to families, the nature of work, household income, disease prevalence, poverty rates, etc. Data may be collected at the national level, county level, district level, and individual unit level. While most of the information sought after may not require to be privacy protected, other information may be sensitive and require the protection of respondent confidentiality/privacy to gain public trust and encourage survey participation. Typically, summary statistics are meant to protect the individuals with sensitive information. However, many state or private agencies may want to carry out detailed statistical analysis beyond just summary statistics, and request access to unit or individual level data. Naturally, releasing such unit level data would most likely compromise the privacy of the respondents and the data collection agencies must address how to protect privacy before releasing them. To address this issue, Statistical Disclosure Avoidance (SDA) methods have been developed and used in most data collection agencies, guaranteeing to the extent possible that the source or identity of the individual providing the sensitive information will be kept confidential. An excellent reference is Drechsler (2011).

Among the available SDA methods, the most popular are (1) noise addition/multiplication, (2) synthetic data based on Plug-In Sampling (PLS), and synthetic data based on posterior predictive sampling (PPS). While methods based on noise addition/multiplication may not require the assumption of a parametric model which generates the observed data, synthetic data analysis methodologies are in general based on an underlying parametric model. Inferential aspects in this case are useful for some meaningful parametric functions [Reiter (2003, 2004, 2005a, 2005b, 2005c), Klein and Sinha (2011, 2013a,b, 2014, 2015, 2016)].

The specific inferential problem we will discuss in this paper is based on count data such as the proportion of units in a random sample possessing an attribute of interest, and our goal is to draw valid inference about the unknown population proportion $\theta$: point estimate, test and confidence interval. If $X \sim Binomial(n, \theta)$, inference about $\theta$ based on $X$ is obvious. What is not obvious is how to draw valid inference about $\theta$ if $X$ cannot be observed, is unavailable, or is simply not released due to privacy considerations! We assume that a perturbed version of $X$, namely, a noise added version and two synthetic versions (PLS/PPS) are available to us. The question then arises how best to use these perturbed versions of $X$ to draw valid inference about $\theta$.

In US Federal Statistical Research Data Center Disclosure Avoidance Methods: A Hand- book for Researchers VERSION 4.0, there is a reference to a paper by Wood et al (2018) which suggests noise addition to X in this context. To be specific, Section A.1 (page 272) of this paper mentions the following approach.

After observing $X = x \sim Binomial(n, \theta)$, we add a noise term $Y$, independent of $X$, and report $Z = x + Y$ . Obviously, $Y$ is assumed to have a mean 0 and some variance. It is suggested that $Y$ be drawn from a normal distribution $N(0, \sigma^2)$ or a Laplace distribution $L(0, \sigma)$ for subsequent inferential purposes. A point estimate of $\theta$ is then taken as $p = Z/n$ with $E(p) = \theta$ and $Var(p) = [\theta(1-\theta)/n + Var(Y)/n^2]$. However, in order for $0 < p < 1$, we must have $0 < Z < n$, which cannot be guaranteed if $X$ and $Y$ are independent. Under either normal or Laplace noise addition, it is quite possible that $-2\sigma < Y < 2\sigma$ which means $X > 2\sigma$ in order that $Z > 0$. Again, for $Z < n$, it must hold that $X + 2\sigma < n$. This should also hold for all possible values of $x$. Naturally a blind use of $Z$ may lead to disastrous results! We address this problem in Section 3 and provide a few acceptable solutions.

## 3. Inference under noise addition

In this section, we identify the steps needed to draw valid inference about $\theta$ based on noise-added data: $Z = X + \varepsilon$, where the noise term $\varepsilon$ is assumed to have mean 0 and a scale parameter $\sigma$. In the sequel, we consider two types of noise distribution: $N(0, \sigma^2)$ and Laplace$(0, \sigma)$. Unlike in Wood et al. (2018) where $Z/n$ is the suggested estimate of $\theta$, we proceed to modify $Z$ in a natural way as $U = max[Z, 0]$ if $Z < 0$ and $U = min\big[[Z], n\big]$ if $Z > n$ where $[Z]$ stands for the integer part of $Z$, which is an obvious choice of $U$ for $0 \le Z \le n$. Note that $Z < 0$ is equivalent to $[\varepsilon/\sigma] < [-X/\sigma]$ and $Z > n$ is the same as $[\varepsilon/\sigma] > [n - X]/\sigma$. We then propose $\hat{\theta} = U/n$. Note that $X$ will not be altered when $|\epsilon/\sigma| < \frac{1}{\sigma}$, implying a privacy protection with $Pr\left[|\epsilon/\sigma| > \frac{1}{\sigma}\right]$. One can choose $\sigma$ appropriately to achieve a desired level of protection. Properties of $\hat{\theta}$ can be easily studied by simulation.

We can also follow a likelihood-based approach based on noise added $Z$ without any further modification to $U$ and propose a maximum likelihood estimate $\hat{\theta}_{mle}(z)$ of $\theta$ based on $Z$. We then compute the observed Fisher information $I_{obs}(\theta)$ and use a standardized version of $\hat{\theta}_{mle}(z)$, namely,

$$U = \left(\hat{\theta}_{\text{mle}}(z) - \theta\right)\sqrt{I_{\text{obs}}\left(\hat{\theta}_{\text{mle}}(z)\right)} \tag{3.1}$$

to test hypotheses about $\theta$ and also to construct a confidence interval for $\theta$. Of course, these results are asymptotic in nature. We use simulations to compute the mean and variance of $\hat{\theta}_{mle}(z)$. We

also included a Bayes estimate of $\theta$ based on a Beta prior $Beta(\alpha, \beta)$, taking $\alpha = \beta = 0.1$, and simulated its mean and variance.

From the point of view of an intruder whose goal is to predict $X$ from the released noise added data $Z$, this will be futile since values of $Z$ often will not make any sense. Next, we consider what a smart intruder would do: predict $X$ based on $Z$ from the conditional distribution of $X$, given $Z$. We note, however, that $Z$ itself is obtained from some known value of $X$, say $x_0$. Once $Z$ is computed, $x_0$ is irrelevant in the subsequent conditional distribution computations. The conditional probability of $X = x_0$, given an observed $Z$, will shed some light on the extent to which privacy about $X$ has been preserved. The unknown parameter $\theta$ appearing in the above conditional distribution can be replaced by $\hat{\theta}_{mle}(z)$ and $\hat{\theta}_{Bayes}(z)$. All details appear in the CSRM Technical Report [Hall et al. (2026)].

Another approach to release sensible count data while protecting privacy can be through the addition of discrete noise terms. Obviously there is no unique method in this context and we have followed the following paradigm based on 3 point shift! After observing $X = x$, we modify it as $Z = x$ with probability $a$, and as $Z = x - 1$ and $Z = x + 1$ each with equal probability $(1 - a)/2$. This is done for all $x = 1, \ldots, n - 1$. At the two extremes $x = 0, n$, we modify it as: For $x = 0$, we define $Z = 0$ with probability $a$ and $Z = 1,2$ each with equal probability $(1 - a)/2$. Likewise, for $x = n$, we define $Z = n$ with probability $a$ and $Z = n - 1, n - 2$ each with equal probability $(1 - a)/2$. In this case (mild) privacy protection probability is $(1 - a)$.

It is rather straightforward to derive the resultant distribution of $Z$ and its mean and variance. Details appear in Section 3.1. It turns out that under this special discrete noise addition scheme, $E(Z/n) \sim E(X/n)$ and also $Var(Z/n) \sim Var(X/n)$, regardless of the value of $a$, implying excellent inference accuracy! Taking $a = 0$ guarantees some privacy protection because the shift is just one point away from the truth! It is quite possible to have more extreme shifts leading to more privacy protection. We have not pursued this here.

### 3.1 Truncated modification

We regularize the noisy binomial count $Z = X + \varepsilon$ by clamping the integer part to the feasible support $\{0,1, \ldots, n\}$:

$$U = \begin{cases} 0, & Z < 0, \\ n, & Z > n, \\ \lfloor Z \rfloor, & \text{otherwise}, \end{cases} \qquad \text{equivalently} \qquad U = \text{clamp}(\lfloor Z \rfloor, 0, n).$$

The estimator $\hat{\theta} = U/n$ is therefore always valid (since $U \in \{0, \ldots, n\}$). Note that while $Z/n$ is unbiased for $\theta$ whenever $\mathbb{E}[\varepsilon] = 0$, the clamping step can introduce finite–sample bias, especially when $\theta$ is close to the boundaries 0 or 1 (i.e., when $U$ is more likely to be truncated at 0 or $n$). Results of simulation study and numercial values of mean, variance and MSE of $U$ are omitted and appear in Hall et al. (2026).

### 3.2 Discrete Noise Addition

Let $P(Z = r \mid X = r) = a$ and $P(X = k) = P_k$ for any $k$. Then, as mentioned earlier, $P(Z = r + 1 \mid X = r) = P(Z = r - 1 \mid X = r) = \frac{1-a}{2}$. We can also show that for all $r \notin \{2, n - 2\}$, we can write

$$P(Z = r) = aP_r + \left(\frac{1-a}{2}\right)[P_{r-1} + P_{r+1}] \tag{3.2}$$

and at the points 2 and $n-2$ we can write

$$P(Z=2) = aP_2 + \left(\tfrac{1-a}{2}\right)[P_0 + P_1 + P_3] \tag{3.3}$$

and

$$P(Z=n-2) = aP(X=n-2) + \left(\tfrac{1-a}{2}\right)[P_{n-3} + P_{n-1} + P_n] \tag{3.4}$$

It is tacitly assumed above that $P_{-1} = P_{n+1} = 0$. We easily verify that the above is a genuine probability distribution of $Z$.

$$\begin{aligned}
\sum_{r=0}^{n} P(Z=r) &= \sum_{r\notin\{2,n-2\}} \left[aP(X=r) + \left(\frac{1-a}{2}\right)(P_{r-1}+P_{r+1})\right] + P(Z=2) + P(Z=n-2) \\
&= \sum_{r=0}^{n} \left[aP_r + \left(\frac{1-a}{2}\right)(P_{r-1}+P_{r+1})\right] + \left(\frac{1-a}{2}\right)[P_0+P_n] \\
&= a + \left(\frac{1-a}{2}\right)\left[\sum_{r=0}^{n} P_{r-1} + \sum_{r=0}^{n} P_r + P_0 + P_n\right] \\
&= a + \left(\frac{1-a}{2}\right)\left[\sum_{r=0}^{n} P_r + \sum_{r=0}^{n} P_r\right] \\
&= a + \left(\frac{1-a}{2}\right)\left[2\sum_{r=0}^{n} P_r\right] \\
&= a + \left(\frac{1-a}{2}\right)2
\end{aligned}$$

The mean and variance of $Z$ can then be readily computed as demonstrated below.

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{r=0}^{n} r\left[aP_r + \left(\frac{1-a}{2}\right)(P_{r-1}+P_{r+1})\right] + 2\left(\frac{1-a}{2}\right)P_0 + (n-2)\left(\frac{1-a}{2}\right)P_n \\
&= a\mathbb{E}[X] + \left(\frac{1-a}{2}\right)([\mathbb{E}[X] - nP_n + (1-P_n)] + [\mathbb{E}[X] - (1-P_0)] + 2P_0 + (n-2)P_n)) \\
&= a\mathbb{E}[X] + \left(\frac{1-a}{2}\right)[2\mathbb{E}[X] + 3(P_0 - P_n)] \\
&= \mathbb{E}[X] + 3\left(\frac{1-a}{2}\right)(P_0 - P_n) \\
&\approx \mathbb{E}[X], \text{ regardless of the value of } a!
\end{aligned}$$

$$\mathbb{E}[Z^2] = a\textstyle\sum_{r=0}^{n} r^2 P_r + \left(\frac{1-a}{2}\right)[\sum_{r=0}^{n} r^2 P_{r-1} + \sum_{r=0}^{n} r^2 P_{r+1} + 4P_0 + (n-2)^2 P_n] \tag{3.5}$$

Note that

$$\begin{aligned}
\sum_{r=0}^{n} r^2 P_{r-1} &= \sum_{u=0}^{n-1} (u+1)^2 P_u \\
&= \sum_{u=0}^{n} (u+1)^2 P_u - (n+1)^2 P_n
\end{aligned}$$

and

$$\sum_{r=0}^{n} r^2 P_{r+1} = \sum_{u=1}^{n} (u-1)^2 P_u$$
$$= \sum_{u=0}^{n} (u-1)^2 P_u - P_0$$

leading to

$$\sum_{r=0}^{n} r^2 P_{r-1} + \sum_{r=0}^{n} r^2 P_{r+1} + 4P_0 + (n-2)^2 P_n$$
$$= 2\mathbb{E}[X^2] + 2 + 3P_0 + ((n-2)^2 - (n+1)^2)P_n$$
$$= 2\mathbb{E}[X^2] + 2 + 3P_0 - 3(2n-1)P_n$$

Hence

$$\mathbb{E}[Z^2] = \mathbb{E}[X^2] + (1-a) + \left(\frac{1-a}{2}\right) 3[P_0 - (2n-1)P_n]$$
$$\approx \mathbb{E}[X^2] + (1-a)$$

We can therefore conclude that $Var[Z] \approx Var[X] + (1-a)$, which implies $Var[Z/n] \approx Var[X/n]$, regardless of $a$. The choice of the constant $a = 0$ will guarantee a drift from the observed data although the drift is just one point away!

## 4. Synthetic Data

An alternative approach for protecting respondent privacy is the creation and publication of "synthetic data" as a substitute for data which might inadvertently reveal sensitive information. Synthetic data is data generated from a model that is intended to capture important characteristics of the original data. There are many potential models that could reasonably be used for this purpose, depending on one's assumptions and the characteristics of the data. This section of the paper focuses on two different models for generating synthetic count data. Section 4.1 will focus on the plug-in sampling (PLS) method, and Section 4.2 focuses on the Bayesian posterior predictive sampling (PPS) method.

### 4.1 Plug-In Sampling (PLS)

The idea behind plug-in sampling (or PLS) is to generate synthetic data from a distribution whose parameters are estimated from the true data. The synthetic data generated by this process are then released instead of the true data, preventing any information about the survey respondents being disclosed.

More concretely, assume as before that we have $X \sim Binomial(n, \theta)$ for some population parameter $\theta$. We can use the point estimate $\hat{\theta} = \frac{1}{n}X$ to generate synthetic data $Z$ from $X$ as

$$Z_{PLS} \sim Binomial(n, \hat{\theta})$$

Unlike data generated via noise addition, synthetic data generated through PLS will always produce non-negative, integer counts. It is straightforward to show based on a standard conditional argument that

$$\mathbb{E}\left[\frac{Z_{PLS}}{n}\right] = \theta$$

and

$$\begin{aligned} Var\left(\frac{Z_{PLS}}{n}\right) &= \frac{1}{n^2}\left(\mathbb{E}[Var(Z_{PLS}|X)] + Var(\mathbb{E}[Z_{PLS}|X])\right) \\ &= \frac{1}{n^2}\left(\mathbb{E}\left[n\left(\frac{X}{n}\right)\left(1-\frac{X}{n}\right)\right]\right) + Var\left(n\left(\frac{X}{n}\right)\right) \\ &= \frac{1}{n^2}\left(\mathbb{E}\left[X-\frac{X^2}{n}\right] + Var(X)\right) \\ &= \left(2-\frac{1}{n}\right)\frac{\theta(1-\theta)}{n} \sim \frac{2\theta(1-\theta)}{n} \end{aligned} \tag{4.1}$$

Specifically, Equation [4.1] shows that when $n$ is reasonably large, the variance of the PLS estimator is almost double the variance of the estimator based on $X$. This additional uncertainty can be interpreted as a "cost" paid by the data user for the privacy guarantee that the PLS synthetic data affords to the survey participants.

One way of quantifying how much privacy is gained due to these protections is to calculate the probability that the synthetic data $Z$ are the same as the protected data $X$. If $Z = X$, then disclosing $Z$ is the same as disclosing $X$ and the PLS procedure does not provide any privacy protections. We can calculate the probability that this occurs as

$$P(Z_{PLS} = x \mid X = x) = \left[\binom{n}{x}\left(\frac{x}{n}\right)^{x}\left(1-\frac{x}{n}\right)^{n-x}\right]$$

Table 4.1 shows what this probability looks like for several values of $n$ and $X/n$. We observe that this probability is substantial when $n$ is small and $X/n$ is far from 0.5, while it is smaller when $n$ increases and $X/n$ moves closer to 0.5. The purpose of using PLS is to provide strong privacy protections for respondent data, and so ideally this probability should be small.

**Table 4.1:** 100 × True Probability $P(Z_{PLS} = x \mid X = x)$ by n and X/n

| n | X/n = 0.1 | X/n = 0.5 | X/n = 0.9 |
|---|---|---|---|
| 10 | 38.74 | 24.61 | 38.74 |
| 20 | 28.52 | 17.62 | 28.52 |
| 40 | 20.59 | 12.54 | 20.59 |
| 60 | 16.93 | 10.26 | 16.93 |
| 80 | 14.71 | 8.89 | 14.71 |
| 100 | 13.19 | 7.96 | 13.19 |

We can also explore what this probability looks like by simulating values for $Z$ for different values of $\theta$, $X/n$, and $n$. Specifically, 10,000 replications of the PLS procedure were generated for each combination $(\theta, n, X/n)$ where $\theta \in \{0.1,0.3,0.5,0.7,0.9\}$, $n \in \{10,20,40,60,80,100\}$, and $X/n \in \{0.1,0.5,0.9\}$. In each set of simulations, the percentage of the time that $Z = X$ in these replications was recorded and documented in Table 4.2. Conditionally on $X/n$, this probability does not depend on the value of $\theta$, and so for simplicity each cell in Table 4.2 represents 50,000 simulation results pooled over $\theta$.

**Table 4.2:** Percentage of PLS Simulations where X is fixed and Z = X by $\theta$ and n

| n | X/n = 0.1 | X/n = 0.5 | X/n = 0.9 |
|---|---|---|---|
| 10 | 38.58 | 24.55 | 38.57 |
| 20 | 28.48 | 17.63 | 28.55 |
| 40 | 20.57 | 12.77 | 20.93 |
| 60 | 16.84 | 10.57 | 17.13 |
| 80 | 14.94 | 8.52 | 14.55 |
| 100 | 13.18 | 7.95 | 13.15 |

The data in Table 4.2 closely resemble the probabilities in Table 4.1. Both Table 4.1 and Table 4.2 suggest that while $\theta$ does not affect $P(Z = X|X)$, the degree of privacy afforded to respondents increases substantially with $n$.

A similar set of simulations in which $X/n$ was not fixed enables us to explore what the marginal probability $P(Z = X)$ looks like. A similar pattern that emerges from the results of these simulations, which are documented in Table 3. As before, $P(Z = X)$ is highest when the sample size $n$ is small. This set of simulations also captures a dependence on $\theta$, however: when $\theta$ is close to 0 or 1, $P(Z = X)$ increases. As $n$ increases and $\theta$ moves closer to 0.5, $P(Z = X)$ decreases to more reasonable values.

**Table 4.3**: Percentage of PLS Simulations where Z = X by $\theta$ and n

| n | $\theta$ = 0.1 | $\theta$ = 0.3 | $\theta$ = 0.5 | $\theta$ = 0.7 | $\theta$ = 0.9 |
|---|---|---|---|---|---|
| 10 | 57.29 | 29.77 | 26.06 | 30.59 | 56.58 |
| 20 | 38.06 | 20.11 | 18.26 | 19.97 | 37.48 |
| 40 | 23.55 | 14.36 | 12.38 | 14.76 | 23.62 |
| 60 | 18.40 | 11.18 | 10.37 | 11.64 | 17.60 |
| 80 | 16.57 | 9.53 | 9.20 | 9.44 | 15.03 |
| 100 | 13.66 | 8.98 | 7.82 | 8.60 | 13.90 |

The same simulations can help us to check the asymptotic distribution of the PLS estimator of $\theta$ at different true values of $\theta$ and sample sizes $n$. In particular, define the standardized $Z$ value for PLS as

$$U_{PLS} = \frac{(Z_{PLS}/n - \theta)}{\sqrt{\frac{2\theta(1-\theta)}{n}}}$$

Figure 1 shows histograms of this $U$ value generated from the same simulations underlying Table 4.3.

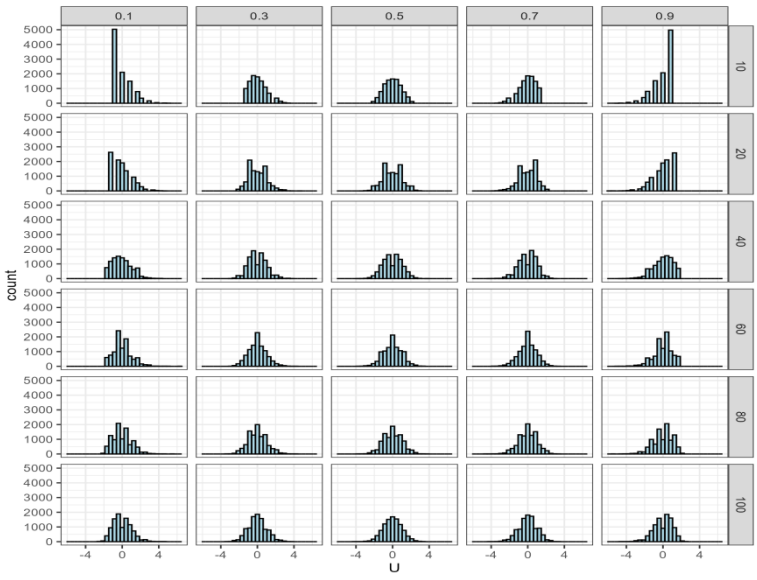


*Figure* 1: $U_{PLS}$ Simulation Histogram by $\theta$ and $n$

Qualitatively, Figure 1 seems to show that the distribution of $U$ approaches normality as $n$ increases, as one might expect. It also seems to show that the speed of this convergence looks faster for $\theta$ values closer to 0.5, and slower for $\theta$ values close to 0 or 1. We can attempt to speed up this convergence by applying a variance stabilizing transformation to the synthetic data estimator. In this case, we may define the variance stabilized version of $U$ as $\widetilde{U}$, and calculate it as

$$\widetilde{U}_{PLS} = \sqrt{2n}\left(arcsin\left(\frac{Z_{PLS}}{n}\right) - arcsin(\theta)\right)$$

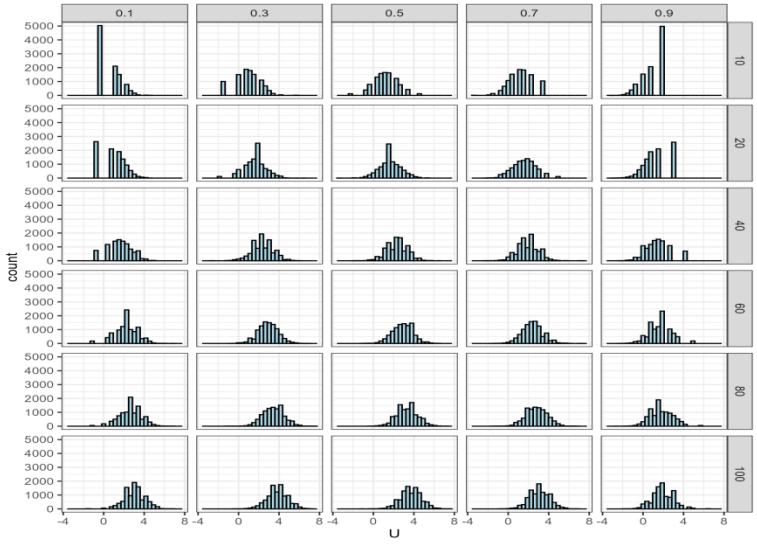


**Figure 2:** Variance Stabilized $\widetilde{U}_{PLS}$ Simulation Histogram by $\theta$ and $n$

The histograms in Figure 2 look slightly more bell-shaped than the non-variance stabilized versions of the simulation for most choices of $\theta$ and $n$, but do not appear to lead to significant improvements in the most difficult settings, e.g. when $\theta$ is the farthest from 0.5 and $n$ is less than or equal to 20.

## 4.2 Posterior Predictive Sampling (PPS)

Posterior predictive sampling (PPS) generates synthetic data based on a Bayesian model. If we specify an appropriate prior distribution for the parameter of interest, we can infer the posterior distribution of each parameter from the data generating function. After this distribution is known or approximated, we draw parameters from their posterior distributions. These parameters are then used to draw synthetic data from its appropriate distribution.

For our binomial problem, we might assign a beta prior to the parameter $\theta$, so that for some hyperparameters $\alpha$ and $\beta$ we have $\theta \sim Beta(\alpha, \beta)$, which combined with $X \sim B(n, \theta)$ yields the posterior $f(\theta|x)$ as $Beta(\alpha + x, \beta + n - x)$.

We can thus generate synthetic data from this model by first drawing $\theta^*$ from

$$\theta^* \sim Beta(\alpha + x, \beta + n - x)$$

and then drawing $Z$ from

$$Z_{PPS} \sim Binomial(n, \theta^*)$$

If this model is correctly specified, then the synthetic data produced by PPS should reflect the characteristics of the original data while protecting respondent privacy. In particular, the expectation of the estimator based on PPS can be easily derived as

$$\begin{aligned}
\mathbb{E}\left[\frac{Z_{PPS}}{n}\right] &= \frac{1}{n}\mathbb{E}\left[\mathbb{E}[Z_{PPS} \mid \theta^*]\right] \\
&= \frac{1}{n}\mathbb{E}[n\theta^*] = \mathbb{E}[\theta^*] \\
&= \mathbb{E}\left[\mathbb{E}[\theta^* \mid X = x]\right] \\
&= \mathbb{E}\left[\frac{\alpha+x}{\alpha+\beta+n}\right] \\
&= \frac{\alpha+n\theta}{\alpha+\beta+n} \approx \theta
\end{aligned} \tag{4.2}$$

When $\alpha = \beta = 0$, this implies that $\frac{1}{n}Z_{PPS}$ is an unbiased estimator of $\theta$. We may still wish to choose non-zero $\alpha$ and $\beta$ values, since this will ensure a proper prior for the PPS mechanism and prevent the posterior distribution $P(Z_{PPS} \mid X)$ from becoming invalid in cases where $X = 0$ or $X = n$. In cases where data producers do not wish for the prior to be informative, it makes sense to choose small values of $\alpha$ and $\beta$ parameters to reduce the bias that this procedure introduces. For this reason, PPS simulations and data tables in this paper will typically assume $\alpha = \beta = 0.01$.

The variance of $Z_{PPS}$ is equal to

$$\begin{aligned}
Var(Z_{PPS}) &= \mathbb{E}[Var(Z_{PPS} \mid \theta^*)] + Var(\mathbb{E}[Z_{PPS} \mid \theta^*]) \\
&= n\left[\mathbb{E}[\theta^*] - \mathbb{E}[(\theta^*)^2]\right] + n^2[\mathbb{E}[(\theta^*)^2] - \mathbb{E}[\theta^*]^2] \\
&= n\mathbb{E}[\theta^*](1 - n\mathbb{E}[\theta^*]) + n(n-1)\mathbb{E}[(\theta^*)^2]
\end{aligned} \tag{4.3}$$

We know that $\mathbb{E}[\theta^*] = \frac{\alpha+n\theta}{\alpha+\beta+n}$. We can calculate $\mathbb{E}[(\theta^*)^2]$ as

$$\begin{aligned}
\mathbb{E}[(\theta^*)^2] &= \mathbb{E}\left[\mathbb{E}[(\theta^*)^2 \mid X = x]\right] \\
&= \mathbb{E}\left[\frac{(x+\alpha)(x+\alpha+1)}{(n+\alpha+\beta)(n+\alpha+\beta+1)}\right] \\
&= \frac{1}{(n+\alpha+\beta)(n+\alpha+\beta+1)}\left(\mathbb{E}[x^2] + (2\alpha+1)\mathbb{E}[x] + \alpha(\alpha+1)\right) \\
&= \frac{\left(n\theta(1-\theta)+n^2\theta^2\right)+(2\alpha+1)n\theta+\alpha(\alpha+1)}{(n+\alpha+\beta)(n+\alpha+\beta+1)} \\
&= \frac{n(n-1)\theta^2+2(\alpha+1)n\theta+\alpha(\alpha+1)}{(n+\alpha+\beta)(n+\alpha+\beta+1)}
\end{aligned} \tag{4.4}$$

Combining Equation [4.4] and Equation [4.2] with Equation [4.3], we can see that

$$Var(Z_{PPS}) = n\left(\frac{\alpha+n\theta}{\alpha+\beta+n}\right)\left(1 - n\left(\frac{\alpha+n\theta}{\alpha+\beta+n}\right)\right) + \tag{4.5}$$

$$n(n-1)\left(\frac{n(n-1)\theta^2+2(\alpha+1)n\theta+\alpha(\alpha+1)}{(n+\alpha+\beta)(n+\alpha+\beta+1)}\right) \tag{4.6}$$

This expression is somewhat complex, but a much simpler approximation can be derived with a bit of work. Here is the final result. Details appear in Hall et al. (2026). First term $T_1$ eventually simplifies to

$$T_1 \approx n\theta - n^2\theta^2 - 2n\theta\alpha + 2(\alpha+\beta)\theta^2 n \tag{4.7}$$

Similarly, we can approximate the second term $T_2$ as

$$T_2 \approx n^2\theta^2 - 2n\theta^2 - 2(\alpha+\beta)n\theta^2 - n\theta^2 + 2n\alpha\theta + 2n\theta \tag{4.8}$$

Combining the approximations for $T_1$ and $T_2$, we arrive at

$$\begin{aligned} Var(Z_{PPS}/n) &= [T_1 + T_2]/n^2 \\ &\approx [n\theta - n^2\theta^2 - 2n\alpha\theta + 2(\alpha+\beta)n\theta^2]/n^2 \\ &+[n^2\theta^2 - 2n\theta^2 - 2(\alpha+\beta)n\theta^2 - n\theta^2 + 2n\alpha\theta + 2n\theta]/n^2 \\ &\approx \tfrac{3\theta(1-\theta)}{n} \end{aligned} \tag{4.9}$$

which is the approximate large-$n$ variance of the PPS synthetic count data based estimate of $\theta$. This is roughly triple the variance of an estimate based on the true data $X$, which suggests a steeper accuracy price is paid by the PPS synthetic data user than the PLS synthetic data user. However, there is some reason to believe that PPS also affords stronger protection to respondents than PLS does.

Professor Gaurisankar Datta (University of Georgia) observed the following connection between PLS and PPS! Comparing the two synthetic data generation schemes, it is clear that if the posterior distribution of $\theta$ under PPS can be made degenerate at just one point $y/n$, then the results under PLS will follow from those under PPS. Generating $\theta^*$ under $Beta\big(k(\alpha+y), k(n-y+\beta)\big)$ for some $k > 0$ and then $Z_{PPS} \sim B(n, \theta^*)$, one can show that

$$Var(Z_{PPS})\ \tfrac{\theta(1-\theta)n^2}{kn+1} + 2n\theta(1-\theta) + O(1/k). \tag{4.10}$$

The above variance expression reduces to $2n\theta(1-\theta)$ for large $k$ (PLS) and to $3n\theta(1-\theta)$ for $k = 1$ (PLS). Obviously, for large $k$, the posterior $Beta\big(k(\alpha+y), k(n-y+\beta)\big)$ reduces to a point mass at $y/n$.

As discussed in Section 4.1 under PLS, here also we can quantify the degree of privacy protection afforded by our PPS synthetic data procedure by evaluating the probability that the synthetic data are the same as the actual data. This probability can be directly computed as a function of $n$, $\alpha$, $\beta$ and $X$ as follows:

$$\begin{aligned} P(Z = x \mid X = x) &= \int_0^1 P\,(Z = x,\, \theta^* \mid X = x)\, d\theta^* = \int_0^1 P\,(Z = x \mid x, \theta^*) \times f(\theta^* \mid x)\, d\theta^* \\ &= \int_0^1 \left[\binom{n}{x} (\theta^*)^x (1-\theta^*)^{n-x}\right] \times \left[\frac{(\theta^*)^{\alpha+x-1}(1-\theta^*)^{\beta+(n-x)-1} d\theta^*}{B\big(\alpha+x, \beta+(n-x)\big)}\right] \\ &= \frac{1}{B\big(\alpha+x, \beta+(n-x)\big)} \binom{n}{x} \int_0^1 (\theta^*)^{\alpha+2x-1}\,(1-\theta^*)^{\beta+2(n-x)-1} d\theta^* \\ &= \binom{n}{x} \frac{B\big(\alpha+2x, \beta+2(n-x)\big)}{B\big(\alpha+x, \beta+(n-x)\big)} \end{aligned} \tag{4.11}$$

Table 4.4 contains some example probability values generated by this function at various values of $n$ and $X/n$, assuming that $\alpha = \beta = 0.01$. Table 4.5 shows the same probabilities as estimated based on 50,000 PPS simulations under the same assumptions. We can see that, in general, these two sets of numbers correspond very closely.

**Table 4.4:** True Probability $P(Z_{PPS} = x \mid X = x)$ by n and X/n

| n | X/n = 0.1 | X/n = 0.5 | X/n = 0.9 |
|---|---|---|---|
| 10 | 26.39 | 17.20 | 26.39 |
| 20 | 19.78 | 12.38 | 19.78 |
| 40 | 14.42 | 8.84 | 14.42 |
| 60 | 11.89 | 7.24 | 11.89 |
| 80 | 10.35 | 6.28 | 10.35 |
| 100 | 9.29 | 5.62 | 9.29 |

**Table 4.5:** Percentage of PPS Simulations where X is fixed and Z = X by n and X/n

| n | X/n = 0.1 | X/n = 0.5 | X/n = 0.9 |
|---|---|---|---|
| 10 | 26.37 | 17.34 | 26.50 |
| 20 | 19.49 | 12.32 | 19.82 |
| 40 | 14.51 | 8.92 | 14.24 |
| 60 | 11.86 | 7.13 | 12.16 |
| 80 | 10.30 | 6.22 | 10.25 |
| 100 | 9.33 | 5.57 | 9.36 |

The same simulations show that the marginal probability of $Z = \theta$ increases as $\theta$ moves away from 0.5, and as $n$ decreases. The same pattern was observed in the PLS simulation table in Section 4.1. Comparing Table 4.3 with Table 4.6, we can observe that the PPS respondent privacy protections appear to be stronger than the PLS respondent privacy protections at all combinations of $n$ and $\theta$, as expressed by the lower estimated probabilities of $Z = X$. This superior privacy protection is the "benefit" derived from the "cost" that the data users pay due to the fact that the variance of the PPS estimator of $\theta$ is higher than the variance of the PLS estimator of $\theta$.

**Table 4.6:** Percentage of PPS Simulations where Z = X by $\theta$ and n

| n | $\theta$ = 0.1 | $\theta$ = 0.3 | $\theta$ = 0.5 | $\theta$ = 0.7 | $\theta$ = 0.9 |
|---|---|---|---|---|---|
| 10 | 50.15 | 21.35 | 18.42 | 21.86 | 50.03 |
| 20 | 30.49 | 14.06 | 12.93 | 13.60 | 30.21 |
| 40 | 16.71 | 9.86 | 8.95 | 10.13 | 16.79 |
| 60 | 12.91 | 7.80 | 7.28 | 8.29 | 12.58 |
| 80 | 10.94 | 6.64 | 6.27 | 6.87 | 11.26 |
| 100 | 9.79 | 5.97 | 5.36 | 6.16 | 9.70 |

The same simulations suggest that the asymptotic distribution of the PPS estimator of $\theta$ also appears to be normal. If we define the standardized $Z$ value for PPS as

$$U_{PPS} = \frac{Z_{PPS}/(n-\theta)}{\sqrt{\frac{3\theta(1-\theta)}{n}}}$$

then Figure 3 shows histograms of $U_{PPS}$ generated from the simulations underlying Table 4.6.

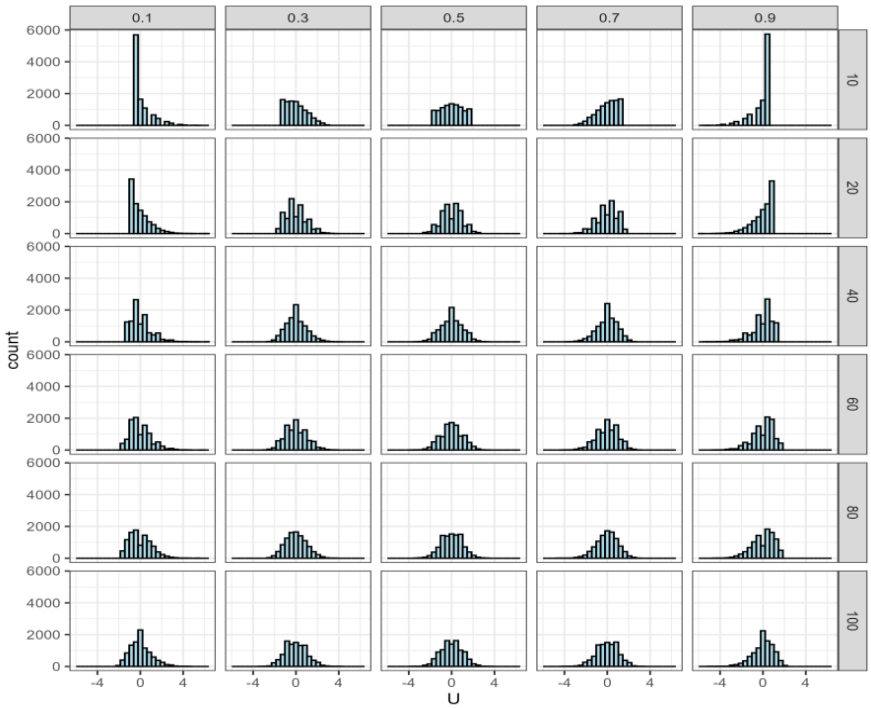


***Figure* 3**: $U_{PPS}$ Simulation Histogram by $\theta$ and $n$

The distributions shown in Figure 3 looks similar to the distributions exhibited by the PLS estimator. As in the PLS synthetic data, values of $\theta$ closer to 0.5 appear to converge to normality faster than values of $\theta$ that are farther from 0.5, and most of the distributions appear to be mostly bell-shaped even for extreme values of $\theta$ when $n$ is at least 40. Subjectively, some of the histograms in Figure 3 appear slightly less bell-shaped than their equivalents in Figure 1 at the same values of $\theta$ and $n$, but the difference is not extreme.

We can try to improve this convergence by applying the following variance stabilizing transformation of $Z_{PPS}$:

$$\widetilde{U}_{PPS} = \sqrt{\frac{n}{3}}\left(2 \times arcsin\left(\frac{Z_{PPS}}{n}\right) - 2 \times arcsin(\theta)\right)$$

Applying this transformation to our simulation data results in the histograms for $\widetilde{U}_{PPS}$ shown in Figure 4.

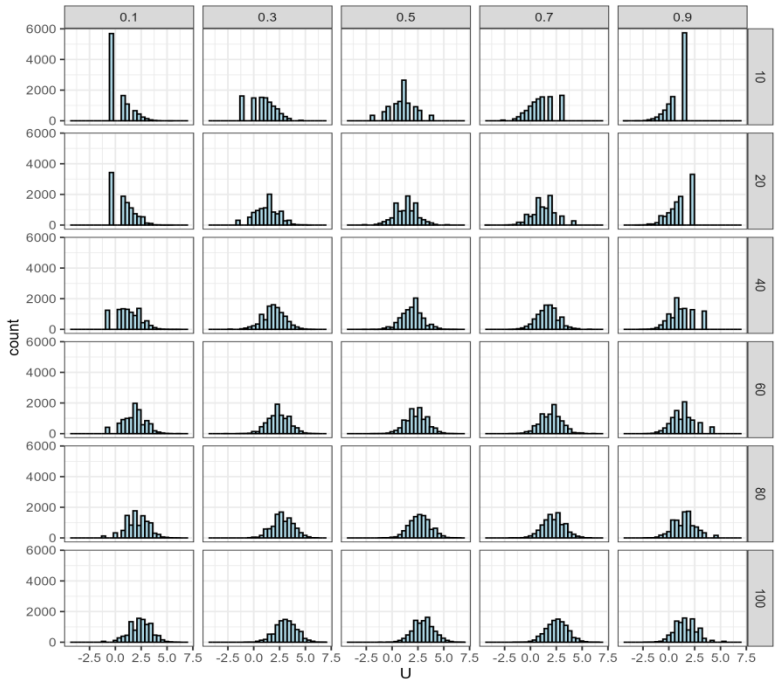


**Figure 4:** Variance Stabilized $\widetilde{U}_{PPS}$ Simulation Histogram by $\theta$ and $n$

As was the case for the variance stabilized $\widetilde{U}_{PLS}$ histograms in Figure 2, this appears to affect the shape of some of the histograms, but does not seem to appreciably speed convergence to normality for the most difficult cases with low $n$ and extreme $\theta$ values.

## 5. Multiple Imputations

Multiple imputations of synthetic data based on PLS and PPS schemes have been widely used based on combination formulas developed by Reiter (2003, 2005, 2007), among others. We now discuss this idea for our three suggested methods.

### 5.1 Noise Addition

It is obvious from our discussion in Section 3 that if we generate multiple imputations of $Z$, namely, $Z_1, \cdots, Z_m$, then for large $m$, the mean of multiply imputed data $\bar{Z}_m$ will converge to $x$, the sensitive value of $X$ used to produce the noise added data, thus violating privacy protection. Although inference accuracy will increase with $m$, we do not recommend this practice here.

### 5.2 Plug-In Sampling (PLS)

Based on our discussion of PLS in Section 4.1, it is again clear that the mean of multiply imputed data $\bar{Z}_m$ based on $Z_1, \cdots, Z_m$, each following Binomial $(n, \hat{\theta})$, will converge to $\hat{\theta} = \frac{x}{n}$ where $x$ is the sensitive value used to generate PLS and meant to be protected. This obviously violates privacy protection and cannot be recommended.

### 5.3 Posterior Predictive Sampling (PPS)

In the context of PPS discussed in Section 4.2, there are two ways we can generate multiple imputations. First, for a fixed random draw $\theta^* \sim Beta(\alpha + x, \beta + n - x)$, we generate multiple $Z$ values, $Z_1, \cdots, Z_m$. By the LLN, $\bar{Z}_m$ will converge to $\theta^*$. Knowing $\theta^*$, the MLE of $x$ can be derived since $n, \alpha, \beta$ are known. We have done some simulations to check to what extent a known value of $X$, $x_0$, used to produce $\theta^*$ is able to reproduce $x_0$. Table 5.1 demonstrates anticipated convergence of $\bar{Z}_m$ to $\theta^*$ while Table 5.2 reproduces $X$ values from PPS-generated $\theta^*$ values for some selected values of $n, X, m, \alpha = \beta = 0.01$.

***Table* 5.1**: $\theta^*$ Estimates based on PPS Imputations with Fixed $\theta^*$

| n | X/n | X | $\theta^*$ | $\bar{\theta}^*$ (m = 5) | $\bar{\theta}^*$ (m = 10) | $\bar{\theta}^*$ (m = 15) | $\bar{\theta}^*$ (m = 20) |
|---|---|---|---|---|---|---|---|
| 10 | 0.10 | 1 | 0.20 | 0.28 | 0.19 | 0.18 | 0.18 |
| 20 | 0.10 | 2 | 0.15 | 0.17 | 0.15 | 0.18 | 0.16 |
| 40 | 0.10 | 4 | 0.13 | 0.10 | 0.15 | 0.15 | 0.11 |
| 60 | 0.10 | 6 | 0.13 | 0.12 | 0.11 | 0.12 | 0.12 |
| 80 | 0.10 | 8 | 0.12 | 0.11 | 0.12 | 0.12 | 0.13 |
| 100 | 0.10 | 10 | 0.12 | 0.12 | 0.12 | 0.10 | 0.13 |
| 10 | 0.50 | 5 | 0.59 | 0.56 | 0.63 | 0.58 | 0.66 |
| 20 | 0.50 | 10 | 0.56 | 0.60 | 0.60 | 0.54 | 0.57 |
| 40 | 0.50 | 20 | 0.55 | 0.53 | 0.55 | 0.54 | 0.54 |
| 60 | 0.50 | 30 | 0.54 | 0.58 | 0.52 | 0.53 | 0.50 |
| 80 | 0.50 | 40 | 0.53 | 0.52 | 0.51 | 0.54 | 0.54 |
| 100 | 0.50 | 50 | 0.53 | 0.54 | 0.52 | 0.54 | 0.54 |
| 10 | 0.90 | 9 | 0.80 | 0.74 | 0.80 | 0.88 | 0.77 |
| 20 | 0.90 | 18 | 0.85 | 0.81 | 0.84 | 0.85 | 0.85 |
| 40 | 0.90 | 36 | 0.87 | 0.88 | 0.87 | 0.88 | 0.88 |
| 60 | 0.90 | 54 | 0.87 | 0.88 | 0.88 | 0.87 | 0.87 |
| 80 | 0.90 | 72 | 0.88 | 0.88 | 0.87 | 0.89 | 0.86 |
| 100 | 0.90 | 90 | 0.88 | 0.90 | 0.88 | 0.87 | 0.88 |

**Table 5.2:** X Estimates based on PPS Imputations with Fixed $\theta^*$

| n | X/n | X | $\theta^*$ | $\hat{X}$ (m = 5) | $\hat{X}$ (m = 10) | $\hat{X}$ (m = 15) | $\hat{X}$ (m = 20) |
|---|---|---|---|---|---|---|---|
| 10 | 0.10 | 1 | 0.20 | 2.80 | 1.90 | 1.80 | 1.80 |
| 20 | 0.10 | 2 | 0.15 | 3.40 | 3.10 | 3.53 | 3.15 |
| 40 | 0.10 | 4 | 0.13 | 3.80 | 6.00 | 5.93 | 4.50 |
| 60 | 0.10 | 6 | 0.13 | 7.20 | 6.70 | 7.20 | 6.95 |
| 80 | 0.10 | 8 | 0.12 | 9.00 | 9.90 | 9.40 | 10.70 |
| 100 | 0.10 | 10 | 0.12 | 11.60 | 12.10 | 10.07 | 12.60 |
| 10 | 0.50 | 5 | 0.59 | 5.60 | 6.30 | 5.80 | 6.60 |
| 20 | 0.50 | 10 | 0.56 | 12.00 | 12.10 | 10.73 | 11.45 |
| 40 | 0.50 | 20 | 0.55 | 21.20 | 21.90 | 21.73 | 21.50 |
| 60 | 0.50 | 30 | 0.54 | 34.60 | 31.20 | 32.00 | 30.30 |
| 80 | 0.50 | 40 | 0.53 | 41.40 | 41.00 | 42.80 | 43.25 |
| 100 | 0.50 | 50 | 0.53 | 54.20 | 52.10 | 53.87 | 54.50 |
| 10 | 0.90 | 9 | 0.80 | 7.40 | 8.00 | 8.80 | 7.70 |
| 20 | 0.90 | 18 | 0.85 | 16.20 | 16.80 | 17.07 | 17.05 |
| 40 | 0.90 | 36 | 0.87 | 35.40 | 34.80 | 35.00 | 35.40 |
| 60 | 0.90 | 54 | 0.87 | 52.60 | 53.10 | 52.27 | 52.10 |
| 80 | 0.90 | 72 | 0.88 | 70.60 | 69.40 | 71.00 | 68.60 |
| 100 | 0.90 | 90 | 0.88 | 90.20 | 88.50 | 87.20 | 88.45 |

Our second method to generate multiple imputations is to independently draw $\theta_1^*, \cdots, \theta_m^*$ from $Beta(\alpha + x, \beta + n - x)$, and then draw one $Z \sim B(n, \theta^*)$ from each selected $\theta^*$. Denoting the resultant $Z$'s by $Z_1, \cdots, Z_m$ and noting from Section 4.2 that each $Z_i/n$ is an unbiased estimate of $\theta$ with $Var(Z_i/n) \sim 3\theta(1-\theta)/n$ and $Cov(Z_i, Z_j) \sim \theta(1-\theta)/n$, a standard meta-analysis method

to combine the $Z_i$'s leads to the following unbiased estimate of $\theta$: $\hat{\theta} = [Z_1 + \ldots \cdots + Z_m]/mn$ with $Var(\hat{\theta}) \sim [3\theta(1-\theta)]/mn$. Here we have used a standard result: If $\mathbf{X} \sim N[\mu\mathbf{1}, \Sigma]$ with $\Sigma$ having an intraclass covariance structure, then the MLE of $\mu$ is $\bar{\mathbf{X}}$. It is clear that the accuracy of inference will increase with $m$. However, as expected, privacy protection will decrease with $m$. Table 5.3 provides such evidence for some values of $m$, taking a rounded value of $\bar{Z}$ as an intruder's obvious prediction of $x$.

**Table 5.3:** Percentage of Simulations with $X = round(\bar{Z})$ based on m PPS Imputations

| n | X/n | % Equal (m = 5) | % Equal (m = 10) | % Equal (m = 15) | % Equal (m = 20) |
|---|---|---|---|---|---|
| 10 | 0.1 | 58 | 67 | 86 | 86 |
| 10 | 0.5 | 40 | 56 | 60 | 69 |
| 10 | 0.9 | 68 | 81 | 90 | 93 |
| 20 | 0.1 | 49 | 67 | 64 | 79 |
| 20 | 0.5 | 31 | 50 | 49 | 60 |
| 20 | 0.9 | 45 | 71 | 72 | 74 |
| 40 | 0.1 | 32 | 46 | 59 | 67 |
| 40 | 0.5 | 27 | 28 | 32 | 44 |
| 40 | 0.9 | 30 | 47 | 50 | 60 |
| 60 | 0.1 | 37 | 46 | 40 | 56 |
| 60 | 0.5 | 14 | 24 | 29 | 37 |
| 60 | 0.9 | 27 | 35 | 38 | 55 |
| 80 | 0.1 | 24 | 34 | 36 | 37 |
| 80 | 0.5 | 17 | 17 | 24 | 24 |
| 80 | 0.9 | 25 | 42 | 38 | 44 |
| 100 | 0.1 | 11 | 32 | 38 | 43 |
| 100 | 0.5 | 12 | 23 | 21 | 22 |
| 100 | 0.9 | 19 | 40 | 36 | 35 |

## 6. Comparison of three methods and concluding remarks

Here we provide a comparison of the three data analysis methods and some concluding remarks.

Under the noise perturbation method, adding normal or Laplace noise makes hardly any difference and, as expected, increasing $n$ increases inferential accuracy while increasing $\sigma$ makes the proposed estimate of $\theta$ less accurate. Also, for large $n$, simulated variance of the proposed estimate of $\theta$ is nearly equal to the reciprocal of observed Fisher information. Moreover, the empirical distribution of the standardized variable $U$ is nearly normal for large sample sizes under both normal and Laplace noise addition, thus making subsequent inference about $\theta$ rather straightforward. In regard to privacy protection, retrieving underlying true value of $x$ from a released data point $z$ becomes increasingly harder as the noise level $\sigma$ increases for any sample size. Thus a judicial choice of $\sigma$ to strike a balance between inference accuracy and privacy protection is warranted. We also note that under the Bayesian approach, the Bayes estimate maintains its unbiasedness just like the MLE with a slightly higher level of variance in most cases. In regard to privacy protection, the performance of the Bayes approach is similar to the use of the MLE-based procedure. Details of the above observations appear in Hall et al. (2026).

Returning to the two synthetic methods of data perturbation, inference about $\theta$ under PLS based on both $U_{PLS}$ and its variance-stabilized version $\bar{U}_{PLS}$ is obvious. Figures 1 and 2 validate the asymptotic normality of $U_{PLS}$ and $\bar{U}_{PLS}$ for large $n$. Tables 4.1-4.3 demonstrate to what extent a specified value of $x$ can be retrieved from the PLS released value $z$ and, as expected, it becomes increasingly difficult with the sample size $n$. Under the PPS scheme, interestingly, the asymptotic

variance of our proposed estimate $Z_{PPS}/n$ of $\theta$ is independent of the two tuning parameters $\alpha$ and $\beta$. Figures 3 and 4 validate the asymptotic normailty of $U_{PPS}$ and $\bar{U}_{PPS}$ for large $n$ and inference based on them is obvious. Tables 4.4-4.6 demonstrate to what extent a specified value of $x$ can be retrieved from the PPS released value $z$ and, as expected, it becomes increasingly difficult with the sample size $n$. Comparing the entries in Tables 4.1-4.3 against those in Tables 4.4 - 4.6, it is obvious that PPS offers more privacy than PLS.

Our last observation relates to the use of multiple imputations under PPS scheme as discussed in Section 5. Inference accuracy increases with the number of imputations $m$ as evident from Table 5.1 and the discussion therein while Tables 5.2 and 5.3 demonstarte that the privacy protection level decreases with increase in $m$.

**Acknowledgement:** Our sincere thanks are due to Professor Bikas Sinha and Professor Gaurisankar Datta for some very helpful comments which led to an improved version of the paper. We very much appreciate Dr. Tommy Wright's encouragement and support.

## References

[1] Drechsler, J., and Reiter, J. P. (2010). Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata. Journal of the American Statistical Association, 105(492), 1347–1357.

[2] Drechsler, J. (2011). Synthetic Datasets for Statistical Disclosure Control. Springer.

[3] Hall, A., Singh, N, and Sinha, B. (2026). Inference about a Binomial Proportion under Privacy Protection. Center for Statistical Research & Methodology, Research Report Series (Statistics #2026-01). U.S. Census Bureau. Available at: https://www.census.gov/library/working-papers/2026/adrm/RRS2026-01.html.

[4] Kim, J. (1986). A method for limiting disclosure in microdata based on random noise and transformation. In Proceedings of the American Statistical Association, Section on Survey Research Methods (pp. 303–308).

[5] Kim, J. J., and Winkler, W. E. (1995). Masking microdata files. In Proceedings of the American Statistical Association, Section on Survey Research Methods (pp. 114–119).

[6] Kim, J. J., and Winkler, W. E. (2003). Multiplicative Noise for Masking Continuous Data. Statistical Research Division, Research Report Series (Statistics #2003-01). U.S. Census Bureau. Available at: http://www.census.gov/srd/papers/pdf/rrs2003-01.pdf.

[7] Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. International Statistical Review, 79(3), 362–384.

[8] Kinney, S. K., Reiter, J. P., and Miranda, J. (2014). SynLBD 2.0: Improving the Synthetic Longitudinal Business Database. Statistical Journal of the IAOS, 30(2), 129–135.

[9] Klein, M., Mathew, T., and Sinha, B. (2013). A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication. Center for Statistical Research & Methodology, Research Report Series (Statistics #2013-02). U.S. Census Bureau. Available at: http://www.census.gov/srd/papers/pdf/rrs2013-02.pdf.

[10] Klein, M., and Sinha, B. (2013). Statistical Analysis of Noise Multiplied Data Using Multiple Imputation. Journal of Official Statistics, 29(3), 425–465.

[11] Klein, M., Mathew, T., and Sinha, B. (2014). Noise Multiplication for Statistical Disclosure Control of Extreme Values in Log-Normal Regression Samples. Journal of Privacy and Confidentiality, 6(1), 77–125.

[12] Klein, M., and Sinha, B. (2015). Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models. Sankhya B, 77(2), 293–311.

[13] Klein, M., and Sinha, B. (2016). Likelihood-Based Finite Sample Inference for Singly Imputed Synthetic Data under the Multivariate Normal and Multiple Linear Regression Models. Journal of Privacy and Confidentiality, 7(1), 43–98.

[14] Lin, Y.-X., and Wise, P. (2012). Estimation of Regression Parameters from Noise Multiplied Data. Journal of Privacy and Confidentiality, 4(1), 61–94.

[15] Little, R. J. A. (1993). Statistical Analysis of Masked Data. Journal of Official Statistics, 9(3), 407–426.

[16] Little, R. J. A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data (2nd ed.). Wiley.

[17] Nayak, T., Sinha, B. K., and Zayatz, L. (2011). Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection. Journal of Official Statistics, 27(3), 527–544.

[18] Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics, 19(1), 1–16.

[19] Reiter, J. P. (2003). Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology, 29(2), 181–188.

[20] Reiter, J. P. (2004). Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. Survey Methodology, 30(2), 235–242.

[21] Reiter, J. P. (2005a). Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. Journal of the Royal Statistical Society, Series A, 168(1), 185–205.

[22] Reiter, J. P. (2005b). Significance Tests for Multi-Component Estimands from Multiply Imputed, Synthetic Microdata. Journal of Statistical Planning and Inference, 131(2), 365–377.

[23] Reiter, J. P. (2005). Using CARTs to Generate Partially Synthetic Public Use Microdata. Journal of Official Statistics, 21(3), 441–462.

[24] Reiter, J. P., and Raghunathan, T. E. (2007). The Multiple Adaptations of Multiple Imputation. Journal of the American Statistical Association, 102(480), 1462–1471.

[25] Reiter, J. P., and Mitra, R. (2009). Estimating Risks of Identification Disclosure in Partially Synthetic Data. Journal of Privacy and Confidentiality, 1(1), 99–110.

[26] Reiter, J. P., and Kinney, S. K. (2012). Inferentially Valid, Partially Synthetic Data: Generating From Posterior Predictive Distributions Not Necessary. Journal of Official Statistics, 28(4), 583–590.

[27] Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley.

[28] Rubin, D. B. (1993). Discussion: Statistical Disclosure Limitation. Journal of Official Statistics, 9(2), 461–468.

[29] Rubin, D. B. (1996). Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91(434), 473–489.

[30] Sinha, B. K., Nayak, T., and Zayatz, L. (2011). Privacy Protection and Quantile Estimation from Noise Multiplied Data. Sankhya B, 73(2), 297–315.

[31] Wang, N., and Robins, J. M. (1998). Large-Sample Theory for Parametric Multiple Imputation Procedures. Biometrika, 85(4), 935–948.

[32] Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D. R., Steinke, T., and Vadhan, S. (2018). Differential Privacy: A Primer for a Non-Technical Audience. Vanderbilt Journal of Entertainment & Technology Law, 1(1), 209–276.