Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions Not Necessary

Jerome P. Reiter* and Satkartar K. Kinney†

1 Introduction

To limit the risks of disclosures when releasing public use data on individual records, statistical agencies and other data disseminators can release multiply-imputed, partially synthetic data (Little, 1993; Reiter, 2003). These comprise the units originally surveyed with some collected values, e.g. sensitive values at high risk of disclosure or values of quasi-identifiers, replaced with multiple imputations. Partially synthetic data can protect confidentiality, since identification of units and their sensitive data can be difficult when select values in the released data are not actual, collected values. And, with appropriate estimation methods based on the concepts of multiple imputation (Rubin, 1987), they enable data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Because of these appealing features, partially synthetic data products have been developed for several major data sources in the U.S., including the Longitudinal Business Database (Kinney et al., 2011), the Survey of Income and Program Participation (Abowd et al., 2006), the American Community Survey group quarters data (Hawala, 2008), and the OnTheMap database of where people live and work (Machanavajjhala et al., 2008). Other examples of partially synthetic data are described in Abowd and Woodcock (2004), Little et al. (2004), Drechsler et al. (2008), and Drechsler and Reiter (2010).

In the statistical theory underlying the generation of partially synthetic data, as well as typical implementations in practice, replacement values are sampled from posterior predictive distributions. That is, the agency repeatedly draws values of the model parameters from their posterior distributions, and generates a set of replacement values based on each parameter draw. The motivation for sampling from posterior predictive distributions derives from multiple imputation of missing data, in which drawing the parameters is necessary to enable approximately unbiased variance estimation (Rubin, 1987, Chapter 4).

In this article, we argue that it is not necessary to draw parameters to enable valid inferences with partially synthetic data. Instead, data disseminators can estimate posterior modes or maximum likelihood estimates of parameters in synthesis models, and simulate replacement values after plugging those modes into the models. Using a simple but informative case, we show mathematically that point and variance estimates based on the plug-in

* Duke University, Box 90251, Durham, NC 27708, USA. Email: jerry@stat.duke.edu
† National Institute of Statistical Sciences, Research Triangle Park, NC 27709, USA. Email: saki@niss.org
method can be approximately unbiased. We also illustrate this fact via simulation studies and include a comparison to generating partially synthetic data from posterior predictive distributions.

The remainder of the article is organized as follows. Section 2 reviews existing methods of generating and making inferences from partially synthetic data. Section 3 offers the mathematical example, and Section 4 presents results of the simulation studies. Section 5 concludes with implications of these results for agencies seeking to generate partially synthetic data.

2 Review of partially synthetic data

To review partially synthetic data, we closely follow the description and notation of Reiter (2003). Let $I_j = 1$ if unit $j$ is selected in the original survey, and $I_j = 0$ otherwise. Let $I = (I_1, \ldots, I_N)$. Let $Y_{\text{obs}}$ be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let $Y_{\text{nobs}}$ be the $(N-n) \times p$ matrix of unobserved survey data for the units with $I_j = 0$; and, let $Y = (Y_{\text{obs}}, Y_{\text{nobs}})$. For simplicity, we assume that all sampled units fully respond to the survey; see Reiter (2004) for simultaneous imputation of missing and synthetic data. Let $X$ be the $N \times d$ matrix of design variables for all $N$ units in the population, e.g., stratum or cluster indicators or size measures. We assume that such design information is known approximately for all population units. It may come, for example, from census records or the sampling frame(s).

The agency releasing synthetic data constructs synthetic data sets based on the observed data, $D = (X, Y_{\text{obs}}, I)$, in a two-part process. First, the agency selects the values from the observed data that will be replaced with imputations. Second, the agency imputes new values to replace those selected values. Let $Z_j = 1$ if unit $j$ is selected to have any of its observed data replaced with synthetic values, and let $Z_j = 0$ for those units with all data left unchanged. Let $Z = (Z_1, \ldots, Z_n)$. Let $Y_{\text{rep},i}$ be all the imputed (replaced) values in the $i$th synthetic data set, and let $Y_{\text{nrep}}$ be all unchanged (unreplaced) values of $Y_{\text{obs}}$. In Reiter (2003), $Y_{\text{rep},i}$ is assumed to be generated from the Bayesian posterior predictive distribution of $(Y_{\text{rep},i}|D, Z)$. The values in $Y_{\text{nrep}}$ are the same in all synthetic data sets. Each synthetic data set, $d_i$, then comprises $(X, Y_{\text{rep},i}, Y_{\text{nrep}}, I, Z)$. Imputations are made independently for $i = 1, \ldots, m$ to yield $m$ different synthetic data sets. These synthetic data sets are released to the public.

Reiter (2003) also describes methods for analyzing the $m$ public use, synthetic data sets. Let $Q$ be the analyst’s scalar estimand of interest, for example the population mean of $Y$ or some coefficient in a regression of $Y$ on $X$. In each $d_i$, the analyst estimates $Q$ with some point estimator $q$ and estimates the variance of $q$ with some estimator $u$. The analyst determines the $q$ and $u$ as if the synthetic data were in fact collected data from a random sample of $(X, Y)$ based on the actual survey design used to generate $I$.

For $i = 1, \ldots, m$, let $q_i$ and $u_i$ be respectively the values of $q$ and $u$ computed with $d_i$. 
The following quantities are needed for inferences:

\[
q_m = \frac{1}{m} \sum_{i=1}^{m} q_i/m \quad (1)
\]

\[
b_m = \frac{1}{m} \sum_{i=1}^{m} (q_i - \bar{q}_m)^2/(m - 1) \quad (2)
\]

\[
\bar{u}_m = \frac{1}{m} \sum_{i=1}^{m} u_i/m \quad (3)
\]

The analyst then can use \(\bar{q}_m\) to estimate \(Q\) and

\[
T_p = b_m/m + \bar{u}_m \quad (4)
\]

to estimate the variance of \(\bar{q}_m\). When \(n\) is large, inferences for scalar \(Q\) can be based on t-distributions with degrees of freedom \(\nu_p = (m - 1)(1 + r_m^{-1})^2\), where \(r_m = (m^{-1}b_m/\bar{u}_m)\). Extensions for multivariate \(Q\) are presented in Reiter (2005a) and Kinney and Reiter (2010).

### 3 Example showing that sampling parameters is unnecessary

In this section, we provide for one scenario a mathematical proof that the estimators \(\bar{q}_m\) and \(T_p\) are approximately unbiased for \(Q\) and the variance of \(\bar{q}_m\), respectively, when generating partially synthetic data without drawing model parameters. For the scenario, we seek to estimate the population mean of a single variable, which we denote \(Y\), in a simple random sample of size \(n\). We do not utilize additional variables for this example; Section 4 displays simulation results involving regressions.

We suppose that the agency replaces all values of \(Y_{obs}\) with draws from some distribution, e.g., all values of \(Y_{obs}\) are confidential. Setting \(Z_j = 1\) for all \(j\) is common in practice; for example, the synthesis for the Longitudinal Business Database, the Survey of Income and Program Participation, and OnTheMap do so. We assume that a reasonable model for the data is \(Y \mid \mu, \sigma^2 \sim N(\mu, \sigma^2)\). Of course, since we have only \(n\) observations in \(Y_{obs}\), we do not know \(\mu\) and \(\sigma^2\). Let \(\bar{y}\) be the sample mean and \(s^2\) be the sample variance, both computed with \(Y_{obs}\). We propose to generate \(m\) partially synthetic data sets with two steps.

**D1.** Sample \(n\) values independently from \(N(\bar{y}, s^2)\), resulting in \(Y_{rep,i}\).

**D2.** Repeat step D1 independently for \(i = 1, \ldots, m\) to create \(m\) partially synthetic data sets that are released to the public.

We note that this process is not sampling from a Bayesian posterior predictive distribution, since we do not draw \((\mu, \sigma^2)\) from their posterior distribution before sampling any \(Y_{rep,i}\).

Using data generated via D1 and D2, in each \(d_i\) we let \(q_i = \bar{y}_i\), i.e., the sample mean in \(d_i\), and let \(u_i = (1 - n/N)s_i^2/n\), where \(s_i^2\) is the usual sample variance of the values in \(d_i\). Hence, we have \(\bar{q}_m = \sum_{i=1}^{m} \bar{y}_i/m; \bar{u}_m = \sum_{i=1}^{m} (1 - n/N)s_i^2/(nm)\); and, \(b_m = \sum_{i=1}^{m} (\bar{y}_i - \bar{y})^2/(m - 1)\).
We next show that \( p \) is unbiased for the actual variance of \( \bar{q}_m \) when averaging over repeated samples \( I \). To begin, we write \( \text{Var}(\bar{q}_m|Y) = E(\text{Var}(\bar{q}_m|Y, I)|Y) + \text{Var}(E(\bar{q}_m|Y, I)|Y) \). From D1, we have

\[
\text{Var}(E(\bar{q}_m|Y, I)|Y) = \text{Var}(\bar{y}|Y) = (1 - n/N)S^2/n, \tag{6}
\]

where \( S^2 = \sum_{i=1}^N (y_i - \bar{Y})^2/(N - 1) \) is the population variance. Also from D1 and D2, we have \( \text{Var}(\bar{q}_m|Y, I) = (s^2/n)/m \), so that

\[
E(\text{Var}(\bar{q}_m|Y, I)|Y) = E(s^2/(nm)|Y) = S^2/(nm). \tag{7}
\]

Hence, we have \( \text{Var}(\bar{q}_m|Y) = S^2/(nm) + (1 - n/N)S^2/n \). Moving to \( E(T_p|Y) \), from D1 we have that \( E(u_i|Y, I) = (1 - n/N)s^2/n \), so that \( E(\bar{u}_m|Y) = (1 - n/N)S^2/n \). Additionally, from D1 we have \( E(b_m|Y, I) = s^2/n \). Hence, we have

\[
E(T_p|Y) = E(\bar{u}_m + b_m/m|Y) = (1 - n/N)S^2/n + S^2/(nm) = \text{Var}(\bar{q}_m|Y). \tag{8}
\]

We note that none of the derivations for the \( t \)-reference distribution in Reiter (2003) require sampling from posterior distributions. Hence, with approximately unbiased point and variance estimates, we can obtain valid variance inferences with these methods.

## 4 Simulation studies

In this section, we illustrate that partial synthesis without posterior predictive simulation can result in well-calibrated inferences. To do so, we generate 10,000 observed data sets \( D \), each comprising \( n = 1000 \) observations and nine variables. For each \( D \), we sample seven of the variables, denoted as \((X_1, \ldots, X_7)\), from independent \( N(0, 1) \). For each observation \( j = 1, \ldots, 1000 \), let \( x_j = (1, x_{j1}, \ldots, x_{j7}) \). For \( j = 1, \ldots, 1000 \), we draw a continuous variable, \( Y_1 \), from the regression \( y_{1j} = x_j'\beta + \epsilon_j \), where \( \beta = (0, -1, 2, -5, 1, 1, 1, 3) \), \( \epsilon_j \sim N(0, \tau^2) \), and \( \tau^2 = 1 \). We also draw a binary variable, \( Y_2 \), using independent Bernoulli distributions such that \( \logit(P(y_{2j} = 1)) = x_j'\alpha + y_{1j}\gamma \). Here, \( \alpha = \beta/3 \) and \( \gamma = -1/3 \). This results in values of \( P(y_{2j} = 1) \) that are between .2 and .8 with high probability. We treat \((Y_1, Y_2)\) as sensitive variables and synthesize all of both. We do not change values of \( X = (X_1, \ldots, X_7) \).

To generate partially synthetic data, we consider two possible strategies. The first is to sample from posterior predictive distributions as recommended in Reiter (2003). We estimate the posterior distributions of \( \beta \) and \( \tau^2 \) based on the default improper prior distribution, \( p(\beta, \tau^2) \propto 1/\tau^2 \). Let \( \hat{\beta} \) be the maximum likelihood estimate (MLE) of \( \beta \), and let \( s^2_{y_{1j}|x} = \sum_{j=1}^n (y_{1j} - x_{j}'\hat{\beta})^2/(n - p) \) be the usual unbiased estimate of \( \tau^2 \). Let \((\hat{\alpha}, \hat{\gamma})\) be the MLE
Table 1: Comparison of simulated coverage rates for 95% confidence intervals and simulated variances of $\bar{q}_m$ when partially synthetic data are created with (Draws) and without (No Draws) sampling from the posterior distributions of the parameters. Results based on 10,000 replications. Variances are reported in parentheses after multiplying by $10^3$.

<table>
<thead>
<tr>
<th></th>
<th>$E(Y_1)$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\beta_5$</th>
<th>$P(Y_1 &gt; 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draws</td>
<td>94.8 (15.6)</td>
<td>94.9 (1.4)</td>
<td>95.4 (1.4)</td>
<td>94.8 (1.4)</td>
<td>94.7 (1.4)</td>
<td>95.2 (1.4)</td>
<td>97.0 (.21)</td>
</tr>
<tr>
<td>No Draws</td>
<td>94.8 (15.4)</td>
<td>94.9 (1.2)</td>
<td>94.8 (1.2)</td>
<td>94.9 (1.2)</td>
<td>94.6 (1.2)</td>
<td>95.2 (1.2)</td>
<td>97.1 (.21)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$E(Y_2)$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
<th>$\alpha_5$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draws</td>
<td>94.9 (.35)</td>
<td>94.6 (12.6)</td>
<td>94.8 (32.0)</td>
<td>95.1 (7.7)</td>
<td>95.2 (6.0)</td>
<td>95.1 (6.0)</td>
<td>94.9 (6.5)</td>
</tr>
<tr>
<td>No Draws</td>
<td>95.1 (.30)</td>
<td>94.6 (10.9)</td>
<td>94.5 (27.5)</td>
<td>94.6 (6.6)</td>
<td>94.7 (5.3)</td>
<td>94.9 (5.3)</td>
<td>94.8 (5.5)</td>
</tr>
</tbody>
</table>

of $(\alpha, \gamma)$, and let $\hat{\Lambda}$ be the estimated covariance matrix of $(\hat{\alpha}, \hat{\gamma})$. These quantities are obtainable from standard logistic regression output. The synthesis process following Reiter (2003) proceeds as follows.

P1. Sample a value of $\tau^2$, say $\tau^{2*}$ from its inverse $\chi^2$ distribution.

P2. Sample a value of $\beta$, say $\beta^*$, from a normal distribution with mean $\hat{\beta}$ and variance $(X'X)^{-1}\tau^{2*}$.

P3. Sample $n = 1000$ values of $Y_1$ from $N(X\beta^*, \tau^{2*})$, resulting in $Y_{1\text{rep},i}$.

P4. Sample a value of $(\alpha, \gamma)$, say $(\alpha^*, \gamma^*)$, from a multivariate normal with mean $(\hat{\alpha}, \hat{\gamma})$ and covariance matrix $\hat{\Lambda}$.

P5. Sample $n = 1000$ values of $Y_2$ from independent Bernoulli distributions such that $\logit(P(y_{2j} = 1)) = x_j'\alpha^* + y_{1\text{rep},i,j}\gamma^*$, resulting in one partially synthetic dataset $(X, Y_{1\text{rep},i}, Y_{2\text{rep},i})$.

P6. Repeat steps P1 to P5 independently $m = 5$ times.

We note that P4 approximates the posterior distribution of $(\alpha, \gamma)$ as a multivariate normal with known covariance. For large $n$, this approximation is reasonable and is typically used in practice.

The second strategy is to sample without drawing parameters. It involves only three steps.

R1. Sample $n = 1000$ values of $Y_1$ from $N(X\hat{\beta}, s^2_{y_{1|x}})$, resulting in $Y_{1\text{rep},i}$.

R2. Sample $n = 1000$ values of $Y_2$ from independent Bernoulli distributions such that $\logit(P(y_{2j} = 1)) = x_j'\hat{\alpha} + y_{1\text{rep},i,j}\hat{\gamma}$, resulting in one partially synthetic dataset $(X, Y_{1\text{rep},i}, Y_{2\text{rep},i})$.

R3. Repeat step R1 to R2 independently $m = 5$ times.
Table 1 displays the simulated coverage rates of 95% confidence intervals, as well as the simulated variances of \( \bar{q}_m \), for the mean of \( Y_1 \), five coefficients in the regression of \( Y_1 \) on \( X \), the percentage of observations with \( Y_1 > 1 \), the mean of \( Y_2 \), and six coefficients in the regression of \( Y_2 \) on \( (Y_1, X) \). The simulated coverage rates in each case are close to the 95% nominal rate, indicating that steps R1 - R3 are sufficient for inferential validity in this simulation. The variances of \( \bar{q}_m \) across the 10,000 replications when data are generated from R1 - R3 are always smaller than those when data are generated from P1 - P6. The magnitude of the variance reduction is minor for the mean of \( Y_1 \) and the \( P(Y_1 > 1) \), but it is generally between 15% and 20% for the other parameters.

We also ran a simulation with \( n = 10,000 \) and otherwise the same design. The 95% confidence interval coverage rates were well-calibrated. The variances of \( \bar{q}_m \) across the 10,000 replications when data were generated from R1 - R3 continued to be always smaller those when data were generated from P1 - P6.

5 Concluding Remarks

Based on the mathematical example and simulations, it appears that agencies do not need to sample from the posterior distributions of parameters to facilitate valid inference from partially synthetic data. This has considerable implications for the generation of partially synthetic data in practice. First, sampling from posterior distributions can be time consuming, as it may require running MCMC algorithms to get posterior distributions. Simply plugging in modes, which often can be computed with off-the-shelf software routines, can reduce this cost. Second, it lends support to the use of synthesizers based on algorithmic methods from machine learning, such as regression trees (Reiter, 2005b), random forests (Caiola and Reiter, 2010), and support vector machines (Drechsler, 2010). These are difficult to justify from the perspective of posterior predictive distributions, since they do not have readily identified model parameters. However, in practice they have been shown to perform reasonably well as data synthesizers (Drechsler and Reiter, 2011). Third, it offers agencies a way to reduce variances of secondary analyses of the released synthetic data.

While synthesizing based on plug-in modes has analytical advantages, it could have disadvantages from the perspective of confidentiality protection. In the setting of Section 3, for example, suppose that an ill-intentioned data snooper knows all values of the variable \( Y \) except for one, say \( y_j \). If the data snooper can get a sharp estimate of \( \bar{y} \) from the synthetic data, he effectively learns the unknown \( y_j \). When synthetic data are generated from \( N(\bar{y}, s^2) \), the data snooper may be able to use \( \bar{q}_m \) and \( \bar{u}_m \) to get close estimates of \( (\bar{y}, s^2) \), and therefore closely estimate the unknown \( y_j \). On the other hand, when synthetic data are generated by drawing \( (\mu, \sigma^2) \) first, the data snooper’s estimate of \( (\bar{y}, s^2) \) has greater uncertainty, and hence his estimate of the unknown \( y_j \) is likely to have higher error. Of course, the “intruder knows all values but one” scenario is an unlikely one in many surveys, and the two approaches may have similar disclosure risk profiles in practice. Nonetheless, the example suggests that evaluating trade-offs in risk and utility from the two partial synthesis strategies is an area for future research.

Many data sets also contain missing values. Reiter (2004) presents an approach to multiple imputation of missing data and synthetic data simultaneously, in which the agency
(i) fills in the missing data by sampling from posterior predictive distributions to create \( m \) completed data sets, and (ii) replaces confidential values in each completed dataset with \( r \) partially synthetic imputations. Hence, a total of \( mr \) nested data sets are released. With this approach, it is necessary to sample from posterior predictive distributions in the first stage of completing the missing values. However, the results in Section 3 and 4 here imply that it is not necessary to use posterior predictive simulation at the second stage.

We also note that it remains necessary to draw from posterior predictive distributions for fully synthetic data (Rubin, 1993; Raghunathan et al., 2003; Si and Reiter, 2011). In fully synthetic data, the agency (i) randomly and independently samples units from the sampling frame to comprise each synthetic data set, (ii) imputes the unknown data values for units in the synthetic samples using models fit with the original survey data, and (iii) releases multiple versions of these data sets to the public. Fully synthetic data essentially involve filling in missing values for records that were not in the original sample. Since one needs to predict values that are not observed, one needs to account for parameter uncertainty in the synthesis models.

References


