Using Synthetic Data Sets to Satisfy Disclosure Restrictions

Jerome P. Reiter
University of California at Santa Barbara

Abstract

Rubin proposed creating multiple, synthetic data sets for public release so that (i) no unit in the released data has sensitive data from an actual unit in the population, and (ii) statistical procedures that are valid for the original data are valid for the released data. Methods for analyzing synthetically created data sets were recently proposed by Raghunathan and Rubin. In this paper, I present results of simulation studies of their methods.

Key Words: Confidentiality, Multiple Imputation, Synthetic Data

1 Introduction

Rubin (1993) proposed creating multiple, synthetic data sets for public release. This approach has three potential benefits. First, it can preserve confidentiality, since identification of units and their sensitive data can be difficult when the data for some or all of the variables in the data set are not actual, collected values. Second, with appropriate estimation methods based on the concepts of multiple imputation (Rubin, 1987), the approach can allow data users to make valid inferences for a variety of estimands without placing undue burdens on these users. Third, synthetic data sets can be sampled by schemes other than the typically complex design used to collect the original data, so that users of synthetic data can ignore the design for inferences.

Methods for analyzing synthetically created data sets were recently proposed by Raghunathan and Rubin (2001). In this paper, I discuss the effectiveness of these methods for making inferences by presenting results of simulation studies.

2 Inferences from Multiple Synthetic Data Sets

To describe construction of and inferences from multiple synthetic data sets, we adapt the notation used for multiple imputation by Rubin (1987). 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_{obs} be the $n \times p$ matrix of collected (real) survey data for the units with $I_j = 1$; let Y_{nobs} be the $(N - n) \times p$ matrix of

unobserved survey data for the units with $I_j = 0$; and, let $Y = (Y_{obs}, Y_{nobs})$. For simplicity, we assume that all sampled units fully respond to the survey. 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 at least approximately, for example from census records or the sampling frames.

The agency releasing synthetic data, henceforth abbreviated as the *imputer*, constructs synthetic data sets based on the observed data (X, Y_{obs}, I) in a two-part process. First, the imputer imputes values of Y for the $N - n$ unobserved units to obtain a completed-data set. The imputer also may choose to impute values of Y for all N units so that the completed-data contains no real values of Y, thereby avoiding the release of any respondent's value of Y. Following Raghunathan and Rubin (2000), we assume that imputations be generated from Bayesian posterior predictive distributions of $(Y|X, Y_{obs}, I)$. Second, the imputer samples units randomly from the completed-data population. These sampled units are released as public use data, so that the released data set contains the values of Y only for units in the synthetic sample. This process is repeated independently m times to get m different synthetic data sets.

We now specify a formal notation for the process of synthetic data construction. Let $(X, Y_{com,i})$ be the completed-data population from which n_{syn} units are sampled to obtain synthetic data set i. Let $Z_{ij} = 1$ if unit j is selected in synthetic data set i, and $Z_{ij} = 0$ otherwise. Let $Z_i = (Z_{i1}, \ldots, Z_{iN})$. Let $Y_{syn,i}$ be the $n_{syn} \times p$ vector of released, synthetic data for units with $Z_{ij} = 1$. The released synthetic data set i is expressed as $(X, Y_{syn,i}, Z_i)$, where all of X is included since design information is assumed known for all units. In practice, it is not necessary to generate completed-data populations for constructing $Y_{syn,i}$. Instead, the imputer need only generate values of Y for units with $Z_{ij} = 1$.

From these synthetic data sets, some user of the publicly released data, henceforth abbreviated as the *analyst*, seeks inferences about some estimand $Q = Q(X, Y)$, where the notation $Q(X, Y)$ means that the estimand Q is a function of (X, Y). For example, Q could be the population mean of Y or the population regression coefficients of Y on X. In each synthetic data set i, the analyst estimates Q with some estimator $Q_i = Q(X, Y_{syn,i}, Z_i)$ and estimates the variance of Q_i with some estimator $V_i = V(X, Y_{syn,i}, Z_i)$. We assume that the analyst determines the Q_i and V_i as if the synthetic data were in fact collected data from a simple random sample of (X, Y).

Under some general conditions on the data imputation process and the estimators used by the analyst, the analyst can obtain randomization valid inferences for Q by combining the Q_i and V_i. Specifically, the following quantities are needed for inferences:

$$\bar{Q}_m = \sum_{i=1}^{m} Q_i/m \tag{1}$$

$$B_m = \sum_{i=1}^{m} (Q_i - \bar{Q}_m)^t (Q_i - \bar{Q}_m)/(m-1) \tag{2}$$

$$\bar{V}_m = \sum_{i=1}^{m} V_i/m. \tag{3}$$

The analyst then can use $\bar{Q}_m$ to estimate Q and

$$T_s = (1 + \frac{1}{m})B_m - \bar{V}_m \tag{4}$$

to estimate the variance of $\bar{Q}_m$. As shown by Raghunathan and Rubin (2001) and Reiter (2001), when the distribution used to draw the synthetic data is the actual posterior predictive distribution of Y, T_s is an unbiased estimator of $Var(\bar{Q}_m)$. When $T_s > 0$, and n, n_{syn}, and m are large, inferences for scalar Q can be based on t-distributions with degrees of freedom

$$v_s = (m-1)(1 - r_m^{-1})^2 \tag{5}$$

where $r_m = (1 + m^{-1})B_m/\bar{V}_m$, so that a $(1-\alpha)\%$ interval for Q is

$$\bar{Q}_m \pm t_{v_s}(\alpha/2)\sqrt{T_s}. \tag{6}$$

Extensions for multivariate Q are not presented here.

Because there may be some estimators for which T_s is negative, particularly when m is modest, it is necessary to have some condition that forces the estimator of $Var(\bar{Q}_m)$ to be positive. Thus, for scalar Q I replace (4) with the modified variance estimator,

$$T_s^* = max(0, T_s) \; + \; \delta \; * (\frac{n_{syn}}{n}\bar{V}_m) \tag{7}$$

where $\delta = 1$ if $T_m < 0$, and $\delta = 0$ otherwise. Appropriate adjustments for the degrees of freedom of the referential t-distribution have not yet been determined. Negative values of T_s generally can be avoided by increasing m.

The variance of $\bar{Q}_m$ in the synthetic data setting differs from the variance of the analogous $\bar{Q}_m$ in the setting of multiple imputation for nonresponse. In the synthetic data setting, the variance calculation involves the distribution used to generate the $(X, Y_{com,i})$ and the random sampling of units from this completed-data population. In the usual multiple imputation setting, the variance calculation involves the distribution used to create imputations for the units with missing data. In fact, as I shall show in simulations, the usual variance formula for multiple imputations, $T_m = (1 + \frac{1}{m})B_m + \bar{V}_m$, tends to overestimate greatly the variance of the synthetic $\bar{Q}_m$.

3 Simulation Studies

I investigate the performance of these methods in simulation studies of four settings:

- estimate a population mean from a simple random sample,
- estimate a regression coefficient from a probability proportional to size sample,
- estimate a regression coefficient from a two-stage cluster sample,
- estimate a population mean from a stratified simple random sample.

The investigations focus on the coverage of asymptotic 95% confidence intervals; they do not examine the potential of the synthetic data approach to preserve confidentiality.

In all simulations, I use the correct posterior predictive distribution to draw synthetic data sets. Of course, in actual implementations, the correct posterior predictive distribution is not known, and an agency-constructed approximation is used. Nonetheless, these idealized simulations help us gauge the promise of releasing synthetic data sets.

3.1 Simple Random Sampling

Assume that we want to estimate the mean of some variable, Y, in a population of size N from a simple random sample of size $n = 100$. Let $Y \sim N(0, 100)$. Further, we'll assume that $N >> n$ so that the finite population correction factor can be ignored when estimating variances.

For each of 500 replications, we construct a collected data set, $Y_{obs} = (Y_1, \ldots, Y_{100})$, by drawing randomly from $Y_j \sim N(0, 100)$ for $j = 1, \ldots, 100$. Using standard noninformative priors on all parameters, the Bayesian posterior predictive distribution of Y,

$$f(Y|Y_{obs}) = \int f(Y|\theta) f(\theta|Y_{obs}) d\theta, \tag{8}$$

where $\theta = (\mu, \sigma^2)$ are the parameters of the normal distribution. To construct a synthetic data set, we use this distribution to draw $n_{syn} = 100$ values of Y_{ij}. This process is repeated independently in $m = 100$ data sets for each replication.

Using the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let Q_i equal the sample mean and V_i equal the sample variance divided by 100. A summary of the actual coverages of 95% confidence intervals for the mean of Y are shown in Table 1. In that table and other tables that follow, "Method" refers to the process of constructing the confidence interval. For example, "Method T_s" means constructing a confidence interval by taking $\bar{Q}_5 \pm t_v \sqrt{T_s}$. The "Observed Data" method uses estimates based on the collected data to form the usual confidence intervals. The column labeled "Avg. $\hat{Q}$" contains the averages across all replications of the point estimates of Q. The column labeled "Avg. $\hat{V}$" contains the averages across all replications of the estimated variances. The column labeled "95% CI cov." contains the percentages of confidence intervals that cover Q.

Table 1: Results for SRS Study (m=100)

Method	Avg. $\hat{Q}$	Avg. $\hat{V}$	95% CI cov.
Observed Data	.04	1.00	94.2
T_s	.04	1.08	93.0
T_m	.04	3.10	100

The average point estimate of the population mean is close to the population value of zero whether we use the actual data or the synthetic data. This is a consequence of using the correct posterior distribution when drawing synthetic data. The actual variance of $\bar{Q}_5$ across the 500 replications is 1.09, verifying that T_s is unbiased. Additionally, $T_s > 0$ in all

500 replications. Confidence coverage is within simulation error of the actual nominal 95% coverage.

The 95% confidence intervals constructed by using MI are too wide. As discussed previously, the distributions used in the development of the variance formulae for multiple imputation differ from the distributions used to create synthetic data sets. In this simulation, these differences make MI an unreliable estimator.

3.2 Probability Proportional to Size Sampling

We now estimate a regression coefficient in a probability proportional to size sample. The hypothetical population is constructed of N=1,000 units, with 4 survey variables, $(X1, X2, X3, X4)$. We draw $X1$ from an exponential distribution, draw $X2 \sim N(0, 3.5)$, draw $X3 \sim N(X1, 3.5)$, and draw $X4 \sim N(X1 + X2 + X3, 100)$. The estimand of interest is the regression coefficient of $X3$ in the regression of $X4$ on $(X1, X2, X3)$, which in the generated population equals 1.07. We assume that $X1$ is known for all units and is available for sampling the collected data and for creating synthetic data sets.

In each of 455 replications, we draw collected data by sampling one hundred units with probability proportional to X1, without replacement, using the scheme of Sunter (1977) as described in Sarndahl, Swensson, and Wretman (1992, pp. 93–96). The ratio of the smallest to largest value of X1 is 42/2, so that the design differs noticeably from simple random sampling.

To create synthetic data, we take a $m = 100$ simple random samples of $n_{syn} = 100$ units from the created population. Since $X1$ is assumed known for all units, we use the actual values of $X1$ for the units in the synthetic data set. To create values of $X2, X3$, and $X4$, we draw from a series of conditional regressions derived from full Bayesian posterior predictive distributions. That is, $X2$ is drawn from its regression on $X1$; $X3$ is drawn from its regression on the synthetically drawn values of $X1$ and $X2$; and, $X4$ is drawn from its regression on the synthetically drawn values of $(X1, X2, X3)$. Standard noninformative priors are assumed for all regression parameters.

Using the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let Q_i equal the estimated regression coefficient of $X3$ in the ordinary least squares regression of $X4$ on $(X1, X2, X3)$, and we let V_i equal the usual estimated variance of this estimated regression coefficient. A summary of the actual coverages of 95% confidence intervals for the regression coefficient is shown in Table 2.

Table 2: Results for PPS Study (m=100)

Method	Avg. $\hat{Q}$	Avg. $\hat{V}$	95% CI cov.
Observed Data	1.15	.29	96.5
T_s	1.15	.30	94.0
T_m	1.15	.91	100

The average point estimates from the observed and synthetic data are close to the value of the population regression coefficient. The actual variance of $\bar{Q}_{100}$ across the 455 replications

is .26, so that T_s appears to be unbiased. T_s is never negative, and approximate 95% nominal coverage is attained. Once again, the multiple imputation variance estimator leads to large overcoverage.

3.3 Two-stage Cluster Sampling

We now estimate a regression coefficient in a two-stage cluster sample. To construct the population, we use the same 1,000 units constructed for the PPS study. To simplify notation, let us call this population POP. Using POP, we form 20 clusters so that each cluster $r = 1, \ldots, 20$ contains the 50 units whose indices are in the range $(50 * (r-1) + 1, 50 * r)$. That is, the first cluster has the first 50 units of POP, the second cluster has the second 50 units of POP, and so on. The values of $X1, X2$, and $X3$ for each unit in this new population are the same values as in POP. For any unit j in cluster r, the value of $X4_{rj}$ is $X4_{rj} = X4_j + \omega_r$, where $X4_j$ is the value of $X4$ in POP and ω_r is a cluster effect. The ω_r are drawn randomly from $\omega_r \sim N(0, 25)$. The estimand of interest is the regression coefficient of $X3$ in the regression of $X4$ on $(X1, X2, X3)$, which in the generated data remains 1.07 after accounting for the clustering.

In 255 replications, we create collected data by sampling in two stages: 1) a simple random sample of 10 clusters; and, 2) within selected clusters, a simple random sample of 10 units. We assume that clustering indicators and $X1$ are known for all units and are available for collecting data and creating synthetic data sets.

To create synthetic data, we take a simple random sample of $n_{syn} = 100$ units from the population. Since $X1$ is assumed known for all units, we can use the values of $X1$ for the units in the synthetic data set. To create $X2$ and $X3$, we draw values from sequential regressions as is done in the PPS simulation. To draw $X4$, we use a three part process. First, we fit a random effects model to the collected data,

$$X4_{rj} = \beta_0 + \beta_1 X1_{rj} + \beta_2 X2_{rj} + \beta_3 X3_{rj} + \omega_r + \epsilon_{rj},$$

where $\epsilon_{rj} \sim N(0, \sigma^2), \omega_r \sim N(0, \tau^2)$. We use this model to determine the posterior distributions of the β's and to find the posterior modes of τ and the ω_r for observed clusters. Second, to estimate ω_r for unobserved clusters, we randomly draw a cluster effect from a normal distribution with mean zero and variance equal to the posterior mode of τ. Finally, we draw a set of β's from their posterior distribution, and then draw new $X4$ from its regression on $(X1, X2, X3)$, conditional on the estimated values of the cluster effects and the drawn values of the β's.

Using the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let Q_i equal the estimated coefficient of the ordinary least squares regression of $X4$ on $(X1, X2, X3)$. We let V_i equal the usual estimated variance of this estimated regression coefficient. A summary of the actual coverages of 95% confidence intervals for the regression coefficient across the 250 replications is shown in Table 4. In that table, the observed data inferences are from the fitted random effects model of $X4$ on $(X1, X2, X3)$, whereas the synthetic data inferences are based on an ordinary least squares model that does not use cluster indicators.

Table 3: Results for CLUS Study (m=100)

Method	Avg. $\hat{Q}$	avg. $\hat{V}$	95% CI cov.
Observed Data	1.05	.33	96.4
T_s	1.04	.32	95.2
T_m	1.04	1.11	100

The average point estimates of the regression coefficient are close to the population value in both the observed and synthetic data. replications is .44. The actual variance of $\bar{Q}_{100}$ across the 250 replications is .27, whowing again that T_s appears to be approximately unbiased. $T_s > 0$ in all replications, and nominal 95% coverage is attained. The observed data inferences based on random effects model, whereas the synthetic data inferences based on regression that ignores clusters.

As in the other studies, using MI leads to overcoverage.

3.4 Stratified Simple Random Sampling

Assume again that we wish to estimate a population mean of some variable, Y. Let each unit j be a member of only one stratum h, where $h = 1, \ldots, 10$ and for all h the size of the stratum N_h=1,000. For all 10,000 units, we construct a population by drawing values from $Y_{hj} \sim N(10 * h, h^2)$. The actual mean of the 10,000 observations in the generated data is 54.94.

Because of the substantial differences in the means and variances across strata, a stratified simple random sample should yield more accurate estimates of the population mean than a simple random sample of the same number of units. That is, the usual unbiased estimator for a stratified random sample, $\bar{y}_{strat} = \frac{N_h}{N} \sum_h \bar{y}_h$, has smaller variance than the usual unbiased estimator for a simple random sample, $\bar{y}$.

In each of 400 replications, we sample a collected data set from this created population by taking a simple random sample of 20 units from each stratum. To construct synthetic data set i, we draw a simple random sample of size $n_{syn} = 200$ from the entire population of 10,000. For sampled synthetic unit j in stratum h, the value of Y_{hj} is drawn from the full Bayesian posterior predictive distribution,

$$f(Y_{hj}|Y_{obs}, H) = \int f(Y|\theta, h) f(\theta_h|Y_{obs}, H) d\theta_h, \tag{9}$$

where $\theta_h = (\mu_h, \sigma_h^2)$ are the parameters of the normal distribution in stratum h, and H is a vector of stratum indicators for all N units. Standard noninformative priors are used for all parameters. This process is repeated in $m = 100$ data sets for each replication.

Using the prescription for analyzing multiple synthetic data sets, in synthetic data set i we let Q_i equal the sample mean and V_i equal the sample variance multiplied by .98 and divided by 200. The .98 is a finite population correction factor. A summary of the actual coverages of 95% confidence intervals for the population mean of Y is shown in Table 4. The observed data inferences are based on the usual unbiased variance estimator for stratified simple random sampling.

Table 4: Results for STRS Study (m=100)

Method	Avg. $\hat{Q}$	Avg. $\hat{V}$	95% CI cov.
Observed Data	54.97	.19	94.3
T_s^*	54.98	1.82	97.2
T_m	54.98	4.32	100.0

The average point estimates from the observed and synthetic data sets are close to the actual population mean. The actual variance across 400 replications of $\bar{Q}_5$ is .22, which is substantially smaller than the averages variance of T^*. This results because $T_s < 0$ for 133 of 400 reps, so that the too large $\bar{V}_{100}$ is used as the variance estimator. We note that synthetic data inferences do not use the stratum indicators, whereas the observed data inferences are based on stratified sampling estimators. If we use the stratum indicators in the synthetic data, then we would be able to improve the estimates of the variances.

4 Simulations using CPS data

At the conference, I plan to present results of simulations using data from the March 2000 Cuurent Population Survey (CPS). These results are not available as of the time of this writing.

5 Concluding Remarks

These simulations, although limited in scope, suggest several conclusions about the use of multiple synthetic data sets for disclosure avoidance. First, with correct posterior predictive distributions, reliable point estimates can be obtained in a variety of settings. Second, the usual multiple imputation formulae result in variance estimates that are too large, which necessitates other methods such as those of Raghunathan and Rubin (2001). Third, with large enough m, it is possible to obtain valid inferences from multiple synthetic data sets for many designs and estimands. In particular, with large m, T_s appears to be a promising method; when $m = 100$ it provides approximate 95% coverage in the SRS, PPS, and 2-stage cluster studies. Fourth, many data sets may need to be released to produce valid inferences in complex sampling designs.

This work is only a beginning step towards understanding the potential of the release of synthetic data for disclosure avoidance. Future research includes the continued development of accurate estimators of variance, assessments of the feasibility of implementing this approach in real data sets, and examinations of whether releasing multiple synthetic data sets does in fact preserve confidentiality.

References

Raghunathan, T. E. and Rubin, D. B. (2001). Multiple imputation for statistical disclosure limitation. *Unpublished technical report* .

Reiter, J. P. (2001). Report on census disclosure project. *Unpublished Census Bureau Technical Report* .

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.