Chapter 7
Partially Synthetic Datasets[1]

As of this writing, no agency has adopted the fully synthetic approach discussed in the previous chapter, but some agencies have adopted a variant of Rubin's original approach suggested by Little (1993): release datasets comprising the units originally surveyed with some collected values, such as sensitive values or values of key identifiers, replaced with multiple imputations. These are called *partially synthetic datasets*. For example, the U.S. Federal Reserve Board protects data in the Survey of Consumer Finances by replacing large monetary values with multiple imputations (Kennickell, 1997). In 2007, the U.S. Census Bureau released a partially synthetic, public use file for the Survey of Income and Program Participation (SIPP) that includes imputed values of social security benefits information and dozens of other highly sensitive variables (`http://www.census.gov/sipp/synth_data.html`). The Census Bureau also protects the identities of people in group quarters (e.g., prisons, shelters) in the public use files of the American Community Survey by replacing demographic data for people at high disclosure risk with imputations. The latest release of a synthetic data product by the Census Bureau is a synthetic version of the Longitudinal Business Database (Kinney et al., 2011) that is available as a public use dataset through the VirtualRDC's Synthetic Data Server located at Cornell University (`http://www.vrdc.cornell.edu/news/data/lbd-synthetic-data/`). Partially synthetic, public use datasets are in the development stage in the U.S. for the Longitudinal Employer–Household Dynamics survey and the American Community Survey veterans and full sample data.

7.1 Inference for partially synthetic datasets

Following Reiter (2003, 2004), let $Z_j = 1$ if unit j is selected to have any of its observed data replaced, and let $Z_j = 0$ otherwise. Let $Z = (Z_1, \ldots, Z_s)$, where s is the number of records in the observed data. Let $Y = (Y_{rep}, Y_{nrep})$ be the data collected

[1] Most of this chapter is taken from Drechsler et al. (2008a) and Drechsler and Reiter (2008).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*,
Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_7,

in the original survey, where Y_{rep} includes all values to be replaced with multiple imputations and Y_{nrep} includes all values not replaced with imputations. Let $Y_{rep}^{(i)}$ be the replacement values for Y_{rep} in synthetic dataset i. Each $Y_{rep}^{(i)}$ is generated by simulating values from the posterior predictive distribution $f(Y_{rep}^{(i)}|Y,Z)$, or some close approximation to the distribution such as those of Raghunathan et al. (2001). The agency repeats the process m times, creating $D^{(i)} = (Y_{nrep}, Y_{rep}^{(i)})$ for $i = 1, \ldots, m$, and releases $\mathbf{D} = \{D^{(1)}, \ldots, D^{(m)}\}$ to the public.

7.1.1 Univariate estimands

To get valid inferences, secondary data users can use the combining rules presented by Reiter (2003). Let Q be an estimand, such as a population mean or regression coefficient. Suppose that, given the original data, the analyst estimates Q with some point estimator q and the variance of q with some estimator u. For simplicity, assume that there are no data with items missing in the observed dataset. Generating partially synthetic datasets if the original data are subject to nonresponse is discussed in Chapter 8. Let $q^{(i)}$ and $u^{(i)}$ be the values of q and u in synthetic dataset $D^{(i)}$ for $i = 1, ..., m$. The analyst computes $q^{(i)}$ and $u^{(i)}$ by acting as if each $D^{(i)}$ is the genuine data. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \tag{7.1}$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \tag{7.2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m. \tag{7.3}$$

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

$$T_p = b_m/m + \bar{u}_m \tag{7.4}$$

to estimate the variance of $\bar{q}_m$.

Similar to the variance estimator for multiple imputation of missing data, b_m/m is the correction factor for the additional variance due to using a finite number of imputations. However, the additional b_m necessary in the missing-data context is not necessary here since $\bar{u}_m$ already captures the variance of Q given the observed data. This is different in the missing-data case, where $\bar{u}_m$ is the variance of Q given the completed data and $\bar{u} + b_m$ is the variance of Q given the observed data.

When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_p = (m-1)(1 + \bar{u}_m/(b_m/m))^2$. Note that the variance estimate

T_p can never be negative, so no adjustments are necessary for partially synthetic datasets.

7.1.2 Multivariate estimands

Significance tests for multicomponent estimands are presented by Reiter (2005c). The derivations are based on the same ideas as those described in Section 5.1.2. Let $\bar{\mathbf{q}}_m$, $\mathbf{b}_m$, and $\bar{\mathbf{u}}_m$ be the multivariate analogs to $\bar{q}_m$, b_m, and $\bar{u}_m$ defined in (7.1) to (7.3). Let us assume the user is interested in testing a null hypothesis of the form $\mathbf{Q} = \mathbf{Q}_0$ for a multivariate estimand with k components. Following the notation in Reiter and Raghunathan (2007), the Wald statistic for this test is given by

$$S_p = (\bar{\mathbf{q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{u}}_m^{-1} (\bar{\mathbf{q}}_m - \mathbf{Q}_0)/(k(1+r_p)), \tag{7.5}$$

where $r_p = (1/m)tr(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_p is an F distribution, F_{k,ν_p}, with $\nu_p = 4 + (t-4)(1+(1-2/t)/r_p)^2$, where $t = k(m-1)$. Synthetic datasets generally require a larger number of imputations m than standard multiple imputation for nonresponse since the fractions of “missing” information tend to be large. Thus, generating less than $m = 4$ synthetic datasets is not recommended, and I do not consider alternative degrees of freedom for $t \leq 4$ as I did in Section 5.1.2.

If $\mathbf{Q}$ contains a large number of components k, using $\bar{\mathbf{u}}_m$ can be cumbersome. As pointed out by Meng and Rubin (1992), it might be more convenient to use a likelihood ratio test in this case. Reiter (2005c) also presents the derivations for this test for partially synthetic datasets.

Again following the notation given in Reiter and Raghunathan (2007), let ψ be the vector of parameters in the analyst's model, and let $\psi^{(i)}$ be the maximum likelihood estimate of ψ computed from $D^{(i)}$, where $D^{(i)}$ is the ith imputed dataset and $i = 1,...,m$. The analyst is interested in testing the hypothesis that $\mathbf{Q}(\psi) = \mathbf{Q}_0$, where $\mathbf{Q}(\psi)$ is a k-dimensional function of ψ. Let $\psi_0^{(i)}$ be the maximum likelihood estimate of ψ obtained from $D^{(i)}$ subject to $\mathbf{Q}(\psi) = \mathbf{Q}_0$. The log-likelihood ratio test statistic associated with $D^{(i)}$ is $L^{(i)} = 2\log f(D^{(i)}|\psi^{(i)}) - 2\log f(D^{(i)}|\psi_0^{(i)})$. Let $\bar{L} = \sum_{i=1}^m L^{(i)}/m$, $\bar{\psi} = \sum_{i=1}^m \psi^{(i)}/m$, and $\bar{\psi}_0 = \sum_{i=1}^m \psi_0^{(i)}/m$. Finally, let $\bar{L}_0 = (1/m)\sum_{i=1}^m (2\log f(D^{(i)}|\bar{\psi}) - 2\log f(D^{(i)}|\bar{\psi}_0))$, the average of the log-likelihood ratio test statistics evaluated at ψ and ψ_0. The likelihood ratio test statistic is given by

$$\hat{S}_p = \bar{L}_0/(k(1+\hat{r}_p)), \tag{7.6}$$

where $\hat{r}_p = (1/t)(\bar{L} - \bar{L}_0)$. The reference distribution for $\hat{S}_p$ is $F_{k,\hat{\nu}_p}$, where $\hat{\nu}_p$ is defined as for ν_p using $\hat{r}_p$ instead of r_p.