Chapter 6
Fully Synthetic Datasets[1]

In 1993, Rubin suggested creating fully synthetic datasets based on the multiple-imputation framework. His idea was to treat all units in the population that have not been selected in the sample as missing data, impute them according to the multiple-imputation approach, and draw simple random samples from these imputed populations for release to the public. Most surveys are conducted using complex sampling designs. Releasing simple random samples simplifies research for the potential user of the data since the design doesn't have to be incorporated in the model. It is not necessary, however, to release simple random samples. If a complex design is used, the analyst accounts for the design in the within-variance $u^{(i)}$, $i = 1, ..., m$.

As an illustration, think of a dataset of size n sampled from a population of size N. Suppose further that the imputer has information about some variables X for the whole population, for example from census records, and only the information from the survey respondents for the remaining variables Y. Let Y_{inc} and Y_{exc} be the observed units and the nonsampled units of the population respectively. For simplicity, assume that there are no data with items missing in the observed dataset. Generating fully synthetic datasets if the original data are subject to nonresponse is discussed in Chapter 8. The synthetic datasets can be generated in two steps. First, construct m imputed synthetic populations by drawing Y_{exc} m times independently from the posterior predictive distribution $f(Y_{exc}|X, Y_{inc})$ for the $N-n$ unobserved values of Y. If the released data should contain no real data for Y, all N values can be drawn from this distribution. Second, take simple random samples from these populations and release them to the public. The second step is necessary, as it might not be feasible to release m whole populations due to the simple matter of data size. In practice, it is not mandatory to generate complete populations. The imputer can make random draws from X in the first step and only impute values of Y for the drawn X. The analysis of the m simulated datasets follows the same lines as the analysis after multiple imputation for missing values in regular datasets, as described in Section 5.1.

[1] Most of this chapter is taken from Drechsler et al. (2008b) and Drechsler and Reiter (2009).

J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, Lecture Notes in Statistics 201, DOI 10.1007/978-1-4614-0326-5_6,

6.1 Inference for fully synthetic datasets

6.1.1 Univariate estimands

To understand the procedure of analyzing fully synthetic datasets, think of an analyst interested in an unknown scalar parameter Q, where Q could be, for example, the mean of a variable, the correlation coefficient between two variables, or a regression coefficient in a linear regression. Inferences for this parameter derived from the original datasets usually are based on a point estimate q, an estimate for the variance of q, u, and a normal or Student's t reference distribution. For analysis of the imputed datasets, let $q^{(i)}$ and $u^{(i)}$ for $i = 1, ..., m$ be the point and variance estimates for each of the m synthetic datasets. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \tag{6.1}$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \tag{6.2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m. \tag{6.3}$$

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

$$T_f = (1 + m^{-1})b_m - \bar{u}_m \tag{6.4}$$

to estimate the variance of $\bar{q}_m$. The difference in this variance estimate compared with the variance estimate for standard multiple imputation (see Section 5.1) is due to the additional sampling from the synthetic units for fully synthetic datasets. Hence, the variance b_m between the datasets already reflects the variance within each imputation. When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_f = (m-1)(1 - \bar{u}_m/((1 + m^{-1})b_m))^2$. Derivations of these methods are presented in Raghunathan et al. (2003).

A disadvantage of this variance estimate is that it can become negative. For that reason, Reiter (2002) suggests a slightly modified variance estimator that is always positive, $T_f^* = \max(0, T_f) + \delta(\frac{n_{syn}}{n}\bar{u}_m)$, where $\delta = 1$ if $T_f < 0$ and $\delta = 0$ otherwise. Here, n_{syn} is the number of observations in the released datasets sampled from the synthetic population.

6.1.2 Multivariate estimands

Significance tests for multicomponent estimands are presented in 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 (6.1) to (6.3). 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

$$S_f = (\bar{\mathbf{q}}_m - \mathbf{Q}_0)^T \bar{\mathbf{u}}_m^{-1} (\bar{\mathbf{q}}_m - \mathbf{Q}_0)/(k(r_f - 1)), \tag{6.5}$$

where $r_f = (1 + 1/m)tr(\mathbf{B}_m \bar{\mathbf{U}}_m^{-1})/k$. The reference distribution for S_f is an F distribution, F_{k,ν_f}, with $\nu_f = 4 + (t-4)(1-(1-2/t)/r_f)^2$, where $t = k(m-1)$. Fully synthetic datasets generally require a larger number of imputations m than standard multiple imputation for nonresponse since the fraction of "missing" information is large (Reiter, 2005b). Thus, generating less than $m = 4$ fully 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 fully 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

$$\hat{S}_f = \bar{L}_0/(k(\hat{r}_f - 1)), \tag{6.6}$$

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

6.2 Analytical validity for fully synthetic datasets

It is important to quantify the analytic usefulness of the synthetic datasets. Existing utility measures are of two types: (i) comparisons of broad differences between the original and released data and (ii) comparisons of differences in specific models between the original and released data. Broad difference measures essentially quantify some statistical distance between the distributions of the original and released data,