ARTICLE IN PRESS

Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Statistical Planning and Inference ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

ELSEVIER

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Significance tests for multi-component estimands from multiply imputed, synthetic microdata

J.P. Reiter*

Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708-0251, USA

Received 1 June 2003; accepted 25 February 2004

Abstract

To limit the risks of disclosures when releasing data to the public, it has been suggested that statistical agencies release multiply imputed, synthetic microdata. For example, the released microdata can be fully synthetic, comprising random samples of units from the sampling frame with simulated values of variables. Or, the released microdata can be partially synthetic, comprising the units originally surveyed with some collected values, e.g. sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations. This article presents inferential methods for synthetic data for multi-component estimands, in particular procedures for Wald and likelihood ratio tests. The performance of the procedures is illustrated with simulation studies.

MSC: 62H15

Keywords: Confidentiality; Disclosure; Multiple imputation; Significance tests; Synthetic data

1. Introduction

When releasing data to the public, statistical agencies seek to provide detailed data while limiting disclosures of respondents' information. Typical strategies for disclosure limitation include recoding variables, swapping data, or adding random noise to data values (Willenborg and de Waal, 2001). However, these methods can distort relationships among variables in the data set. They also complicate analyses for users: to

* Tel.: +1-916-6685227; fax: +1-919-6848594.
E-mail address: jerry@stat.duke.edu (J.P. Reiter).

0378-3758/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.jspi.2004.02.003

analyze perturbed data properly, users should apply the likelihood-based methods described by Little (1993) or the measurement error models described by Fuller (1993). These are difficult to use for non-standard estimands and may require analysts to learn new statistical methods and specialized software programs.

An alternative approach is to release multiply imputed, synthetic microdata. This approach was first suggested by Rubin (1993), who proposed that agencies (i) randomly and independently sample units from the sampling frame to comprise each synthetic data set, and (ii) impute unknown data values for units in the synthetic samples using models fit with the original survey data. These are called *fully synthetic* data sets. Inferences for scalar estimands from fully synthetic data sets can be made using the methods developed by Raghunathan et al. (2003), whose rules for combining point and variance estimates differ from the rules for multiple imputation of missing data (Rubin, 1987). Other authors suggest releasing data sets comprising the units originally surveyed with some collected values, such as sensitive values at high risk of disclosure or values of key identifiers, replaced with multiple imputations (Little, 1993; Kennickell, 1997; Abowd and Woodcock, 2001; Liu and Little, 2002). These are called *partially synthetic* data sets. Inferences for scalar estimands from partially synthetic data sets can be made using the methods developed by Reiter (2003), whose rules for combining point and variance estimates again differ from those of Rubin (1987) and also from those of Raghunathan et al. (2003). Other variants and discussions of synthetic data approaches appear in Fienberg et al. (1996, 1998), Dandekar et al. (2002a, b), Franconi and Stander (2002, 2003), Polettini et al. (2002), Polettini (2003), and Reiter (2002, 2004).

Releasing synthetic data is an appealing approach to disclosure limitation. It can protect confidentiality, since identification of units and their sensitive data can be difficult when some or all of the values in the released data are not actual, collected values. And, with appropriate estimation methods based on the concepts of multiple imputation (Rubin, 1987), it can allow data users to make valid inferences for a variety of estimands using standard, complete-data statistical methods and software. Other attractive features of synthetic data are described by Rubin (1993), Little (1993), Raghunathan et al. (2003), and Reiter (2003, 2004).

This paper presents methods for testing hypotheses about multi-component estimands from multiply imputed synthetic data, specifically procedures for performing Wald and likelihood ratio tests of multi-component null hypotheses. This extends the work of Raghunathan et al. (2003) and Reiter (2003), who develop inferential methods for scalar estimands but do not consider multi-component hypothesis testing. One particularly appealing feature of the likelihood ratio tests developed here is that users can test multi-component hypotheses without estimated covariance matrices, i.e. they just need to evaluate functions of the likelihood. This feature is lacking in simple multivariate generalizations of the formulas of Raghunathan et al. (2003) and Reiter (2003).

The procedures are based on the theory of significance testing for multiple imputation for missing data (Li, 1985; Raghunathan, 1987; Rubin, 1987; Li et al., 1991; Meng and Rubin, 1992), but the test statistics and reference distributions for the synthetic data procedures differ from their multiple imputation counterparts. The procedures for both fully and partially synthetic data are presented in Section 2 and derived in Section

3. Section 4 describes results of simulation studies that illustrate how these procedures can have near nominal significance levels.

2. Significance tests from synthetic data

In this paper, we index quantities for fully synthetic data with a subscript "f" and quantities for partially synthetic data with a subscript "p." Suppose m synthetic data sets are released. It is assumed that imputations are drawn from appropriate posterior predictive distributions, conditional on the observed data. Details on generating fully and partially synthetic data are provided in Raghunathan et al. (2003) and Reiter (2003), respectively.

Using these m data sets, some analyst seeks to test the null hypothesis $Q = Q_0$ for some k-component estimand Q, for example to test if k regression coefficients equal zero. In each synthetic data set d_i, for $i = 1, \ldots, m$, the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator v. It is assumed that the analyst determines the q and v as if (i) for fully synthetic data, the d_i are simple random samples from the sampling frame (Raghunathan et al., 2003), and (ii) for partially synthetic data, the d_i are samples from the sampling frame taken by the original sampling design (Reiter, 2003).

For $i = 1, \ldots, m$, let q_i and v_i be, respectively, the values of q and v in synthetic data set d_i. The following multivariate quantities are needed for inferences for multi-component Q:

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

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

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

The $\bar{q}_m$ is the average of the synthetic point estimates; the b_m is the covariance matrix of these point estimates; and, the $\bar{v}_m$ is the average across synthetic data sets of the estimated covariance matrices. These quantities are needed for both fully and partially synthetic data inferences.

2.1. Wald tests from synthetic data

When the entire $k \times k$ covariance matrix $\bar{v}_m$ is available, analysts can use Wald-type test statistics to test $Q = Q_0$. The test statistics for fully and partially synthetic data are, respectively,

$$S_{\mathrm{f}} = (\bar{q}_m - Q_0)^t \bar{v}_m^{-1} (\bar{q}_m - Q_0)/[k(r_{\mathrm{f}} - 1)], \tag{4}$$

$$S_{\mathrm{p}} = (\bar{q}_m - Q_0)^t \bar{v}_m^{-1} (\bar{q}_m - Q_0)/[k(1 + r_{\mathrm{p}})], \tag{5}$$

where

$$r_f = (1 + 1/m)\,\mathrm{trace}(b_m \bar{v}_m^{-1})/k, \tag{6}$$

$$r_p = (1/m)\,\mathrm{trace}(b_m \bar{v}_m^{-1})/k. \tag{7}$$

The test statistics in (4) and (5) have the familiar quadratic forms of Wald statistics, but additionally there are correction factors $r_f - 1$ or $1 + r_p$ in the denominators. To motivate the need for these correction factors, and to see why they differ for the two types of synthetic data, it is instructive to consider the case where Q is scalar ($k = 1$). For fully synthetic data, Raghunathan et al. (2003) show that a reasonable estimate of the posterior variance of scalar Q is $T_f = (1 + 1/m)b_m - \bar{v}_m$. Hence, for scalar Q, the $r_f - 1 = T_f/\bar{v}_m$, so that $r_f - 1$ estimates the increase in the variance of Q due to using fully synthetic data, relative to $\bar{v}_m$. Similarly, for partially synthetic data, Reiter (2003) shows that $T_p = (1/m)b_m + \bar{v}_m$ is a reasonable posterior variance estimate for scalar Q, so that $1 + r_p = T_p/\bar{v}_m$. When Q is multivariate, the $r_f - 1$ and $1 + r_p$ can be interpreted as the average relative increases in variance across the components of Q. From these interpretations, we can see that the $r_f - 1$ and $1 + r_p$ adjust the quadratic form in (4) and (5) so that the test statistic is based on an appropriate estimate of the variance of Q.

Reference distributions for S_f and S_p are approximated by F-distributions, $F_{k,w(r_f)}$ and $F_{k,w(r_p)}$, with

$$w(r_f) = 4 + (t - 4)(1 - (1 - 2/t)/r_f)^2, \tag{8}$$

$$w(r_p) = 4 + (t - 4)(1 + (1 - 2/t)/r_p)^2, \tag{9}$$

where $t = k(m - 1)$. The p-value for testing $Q = Q_0$ in fully synthetic data is $P_f = \Pr(F_{k,w(r_f)} > S_f)$, and the p-value for testing $Q = Q_0$ in partially synthetic data is $P_p = \Pr(F_{k,w(r_p)} > S_p)$. The $w(r_f)$ and $w(r_p)$ differ because of differences in the Taylor series approximations used to derive the reference distributions.

2.2. Log-likelihood ratio tests from synthetic data

As observed by Meng and Rubin (1992), it may be cumbersome to work with $\bar{v}_m$ for large k, and some software packages do not make readily available covariance matrices of parameter estimates. It is desirable to develop alternative significance tests based on the sets of log-likelihood ratio test statistics from the m synthetic data sets, which are easily computed for common models like those from exponential families.

These tests are derived using the strategy of Meng and Rubin (1992), in which we (i) find a statistic asymptotically equivalent to S_f based only on values of the Wald statistics calculated in each synthetic data set; (ii) use the asymptotic equivalence of Wald and likelihood ratio test statistics to define the likelihood ratio test statistic; and, (iii) use a reference F distribution like the one for the Wald tests. The key to this strategy is to approximate S_f and S_p, as well as $w(r_f)$ and $w(r_p)$, without using $\bar{v}_m$.

Following the notation of Schafer (2000), let ψ be the vector of parameters in the analyst's model, and let $\hat{\psi}_i$ be the maximum likelihood estimate of ψ obtained from d_i. Suppose the user is interested in a k-dimensional function, $Q(\psi)$, and forms the hypothesis that $Q(\psi) = Q_0$. Let $\hat{\psi}_{0i}$ be the maximum likelihood estimate of ψ obtained from d_i subject to $Q(\psi) = Q_0$. The log-likelihood ratio test statistic associated with d_i is

$$l_i = 2 \log f(d_i \mid \hat{\psi}_i) - 2 \log f(d_i \mid \hat{\psi}_{0i}) \tag{10}$$

and their average is $\bar{l} = \sum_{i=1}^{m} l_i/m$. Let the averages of the maximum likelihood estimates be $\bar{\psi} = \sum_{i=1}^{m} \hat{\psi}_i/m$ and $\bar{\psi}_0 = \sum_{i=1}^{m} \hat{\psi}_{0i}/m$. Following Meng and Rubin (1992), we also use the average of the log-likelihood ratio test statistics evaluated at $\bar{\psi}$ and $\bar{\psi}_0$:

$$\bar{L} = (1/m) \sum_{i=1}^{m} (2 \log f(d_i \mid \bar{\psi}) - 2 \log f(d_i \mid \bar{\psi}_0)). \tag{11}$$

The likelihood ratio test statistics for fully and partially synthetic data are

$$\hat{S}_{\mathrm{f}} = \bar{L}/(k(\hat{r}_{\mathrm{f}} - 1)), \tag{12}$$

$$\hat{S}_{\mathrm{p}} = \bar{L}/(k(\hat{r}_{\mathrm{p}} + 1)), \tag{13}$$

where $\hat{r}_{\mathrm{f}} = ((m+1)/t)(\bar{l} - \bar{L})$ and $\hat{r}_{\mathrm{p}} = (1/t)(\bar{l} - \bar{L})$.

In (12) and (13), the $\bar{L}$ essentially is an asymptotically equivalent replacement for the quadratic form in the Wald statistics in (4) and (5). The $\hat{r}_{\mathrm{f}}$ and $\hat{r}_{\mathrm{p}}$ replace r_{f} and r_{p}, respectively, so that $\hat{r}_{\mathrm{f}} - 1$ and $1 + \hat{r}_{\mathrm{p}}$ adjust for the relative increase in variance due to using synthetic data. Importantly, these new quantities do not require knowledge of $\bar{v}_m$ to make the adjustments.

The reference distributions for $\hat{S}_{\mathrm{f}}$ and $\hat{S}_{\mathrm{p}}$ are $F_{k,w(\hat{r}_{\mathrm{f}})}$ and $F_{k,w(\hat{r}_{\mathrm{p}})}$, where the $w(\hat{r}_{\mathrm{f}})$ and $w(\hat{r}_{\mathrm{p}})$ are defined as in (8) and (9) using $\hat{r}_{\mathrm{f}}$ and $\hat{r}_{\mathrm{p}}$.

3. Derivation of the tests

This section shows derivations of the tests in Section 2. The derivations are based on the theory for large sample significance tests in multiple imputation for missing data (Li, 1985; Raghunathan, 1987; Rubin, 1987; Li et al., 1991; Meng and Rubin, 1992). For both fully and partially synthetic data settings, let $d^m = \{d_1, d_2, \ldots, d_m\}$ represent the collection of released synthetic data sets.

3.1. Tests for fully synthetic data

It is helpful to conceive of the generation of each fully synthetic data set d_i as a two step process, as done in Raghunathan et al. (2003). First, the imputer imputes values for all units not in the original survey, thereby creating a completed-population

D_i. Then, the imputer takes a simple random sample from D_i to obtain d_i. Let Q_i be the completed-population value of Q obtained from D_i, and let

$$\bar{Q}_m = \sum_{i=1}^{m} Q_i/m, \tag{14}$$

$$B_m = \sum_{i=1}^{m} (Q_i - \bar{Q}_m)(Q_i - \bar{Q}_m)^t/(m-1). \tag{15}$$

Extending the Bayesian theory of Raghunathan et al. (2003) for scalar Q to multivariate Q, we obtain

$$(Q \mid \bar{Q}_m, B_m) \sim t_{m-1}(\bar{Q}_m, (1+1/m)B_m), \tag{16}$$

$$(\bar{Q}_m \mid d^m) \sim \mathrm{N}(\bar{q}_m, \bar{v}_m/m), \tag{17}$$

$$((m-1)b_m(B_m + \bar{v}_m)^{-1} \mid d^m) \sim \mathrm{Wishart}_{m-1}. \tag{18}$$

The data do not need to be multivariate normal for this theory to hold; rather, the posterior distribution of the population estimand Q needs to be normal (for large m). This is reasonable for many common estimands, including population means and regression coefficients, across a wide variety of data distributions.

Posterior inferences for multivariate $(Q \mid d^m)$ are derived by integrating the product of (16)–(18) with respect to $\bar{Q}_m$ and B_m. The integral can be simplified considerably by replacing (16) with

$$(Q \mid \bar{Q}_m, B_m) \sim \mathrm{N}(\bar{Q}_m, (1+1/m)B_m), \tag{19}$$

so that

$$(Q \mid B_m, d^m) \sim \mathrm{N}(\bar{q}_m, (1+1/m)B_m + \bar{v}_m/m). \tag{20}$$

This simplification should be reasonable for relatively large m.

Conditional on B_m, a p-value for a significance test of $Q = Q_0$ is obtained from the Wald test statistic associated with (20) and a χ^2_k reference distribution. However, the analyst does not know B_m and so must integrate the expression for the p-value, P_f, with respect to B_m. The resulting integral is

$$P_\mathrm{f} = \int \Pr(\chi^2_k > (\bar{q}_m - Q_0)^t((1+1/m)B_m + \bar{v}_m/m)^{-1}(\bar{q}_m - Q_0)) \times f(B_m \mid d^m)\,\mathrm{d}B_m. \tag{21}$$

where χ^2_k is a chi-squared random variable on k degrees of freedom. The integral in (21) can be evaluated numerically, but it is desirable to have a closed-form approximation. Following Rubin (1987) and Li et al. (1991), we assume that B_m is proportional to $\bar{v}_m$, i.e. $B_m = R_\mathrm{f}\bar{v}_m$ where R_f is a scalar, so that

$$P_\mathrm{f} = \int \Pr(\chi^2_k > (\bar{q}_m - Q_0)^t\bar{v}_m^{-1}(\bar{q}_m - Q_0)((1+1/m)(1+R_\mathrm{f}) - 1)^{-1}) \times f(R_\mathrm{f} \mid d^m)\,\mathrm{d}R_\mathrm{f}. \tag{22}$$

The proportionality assumption may be reasonable in many fully synthetic data settings, because all variables have 100% simulated values.

Using (18) and standard multivariate normal theory

$$(k(m-1)(\text{trace}(b_m\bar{v}_m^{-1})/k)/(1+R_f)\,|\,d^m) \sim \chi^2_{k(m-1)}. \tag{23}$$

Substituting (23) into (22), and using r_f as defined in (6) and $t = k(m-1)$, we obtain

$$P_f = \Pr(\chi_k^2 > (\bar{q}_m - Q_0)^t \bar{v}_m^{-1}(\bar{q}_m - Q_0)(r_f t/\chi_t^2 - 1)^{-1}) \tag{24}$$

$$= \Pr((\chi_k^2/k)[(r_f t/\chi_t^2 - 1)/(r_f - 1)] > S_f). \tag{25}$$

Following Li et al. (1991), we approximate the distribution of $(\chi_k^2/k)[(r_f t/\chi_t^2 - 1)/(r_f - 1)]$ as a multiple of an F distribution, $\delta F_{k,w}$. The approximation matches the first two moments of $F_{k,w}$ with the first two moments of the Taylor series expansion of $(\chi_k^2/k)[(r_f t/\chi_t^2 - 1)/(r_f - 1)]$. The series is expanded in $1/\chi_t^2$ around its expectation $1/(t-2)$ as follows:

$$\begin{aligned}\delta F_{k,w} \approx{}& (\chi_k^2/k)[(r_f t/(t-2) - 1)/(r_f - 1)] \\ &+ (\chi_k^2/k)[r_f t/(r_f - 1)](1/\chi_t^2 - 1/(t-2)),\end{aligned} \tag{26}$$

$$E(\delta F_{k,w}) = \delta w/(w-2) \approx (r_f t/(t-2) - 1)/(r_f - 1), \tag{27}$$

$$\begin{aligned}E(\delta^2 F_{k,w}^2) &= \delta^2 (w/k)^2 k(k+2)/[(w-2)(w-4)] \\ &\approx (k(k+2)/k^2)([(r_f t/(t-2) - 1)/(r_f - 1)]^2 \\ &\quad + [r_f t/(r_f - 1)]^2 (2/[(t-2)^2(t-4)])).\end{aligned} \tag{28}$$

Solving, we obtain $\delta = (1 - 2/w)(r_f t/(t-2) - 1)/(r_f - 1)$, and $w = w(r_f)$ as defined in (8). Setting $\delta = 1$, as it will be approximately for large t, results in $F_{k,w(r_f)}$ of Section 2.

Likelihood ratio tests can be derived using the strategy in Meng and Rubin (1992), namely (i) find a statistic asymptotically equivalent to S_f based only on values of the Wald statistics from each synthetic data set; (ii) use the asymptotic equivalence of Wald and likelihood ratio test statistics to define the likelihood ratio test statistic; and, (iii) use a reference F distribution like the one for the Wald tests.

To take the first step, we require two different averages of the synthetic data Wald statistics

$$\bar{w} = (1/m)\sum_{i=1}^{m} (q_i - Q_0)^t v_i^{-1}(q_i - Q_0), \tag{29}$$

$$\bar{W} = (1/m)\sum_{i=1}^{m} (\bar{q}_m - Q_0)^t v_i^{-1}(\bar{q}_m - Q_0). \tag{30}$$

Using arguments like those in Rubin (1987, pp. 99–100), a statistic asymptotically equivalent to S_f is

$$(\bar{w}/k - r_f(m-1)/(m+1))/(r_f - 1). \tag{31}$$

To verify this, we assume without loss of generality (Rubin, 1987, p. 100) that $Q_0 = 0$ and $\bar{v}_m^{-1}$ is a $k \times k$ identity matrix. Then,

$$S_f = (r_f - 1)^{-1} \bar{q}_m \bar{q}_m^t / k \tag{32}$$

and

$$\begin{aligned} \bar{w} &= \sum_{i=1}^{m} q_i q_i^t / m = \bar{q}_m \bar{q}_m^t + \sum_{i=1}^{m} (q_i - \bar{q}_m)(q_i - \bar{q}_m)^t / m \\ &= \bar{q}_m \bar{q}_m^t + (m-1)\text{trace}(B_m)/m \\ &= \bar{q}_m \bar{q}_m^t + (m-1) k r_f / (m+1). \end{aligned} \tag{33}$$

Substituting (33) for $\bar{w}$ into (31) obtains (32).

The r_f defined in (6) requires $\text{trace}(b_m \bar{v}_m^{-1})$, which we do not want to use when deriving these tests. An expression for r_f that relies only on Wald statistics is obtained by setting (31) equal to (4), resulting in

$$\tilde{r}_f = (m+1)(1/t)(\bar{w} - \bar{W}). \tag{34}$$

Here, $\bar{W}$ is used to approximate $(\bar{q}_m - Q_0)^t \bar{v}_m^{-1} (\bar{q}_m - Q_0)$ in (4). Using $\bar{W}$ and $\tilde{r}_f$, we can re-express S_f as the asymptotically equivalent

$$\tilde{S}_f = \bar{W} / k(\tilde{r}_f - 1). \tag{35}$$

Taking step 2 of the Meng and Rubin (1992) strategy, we use the asymptotic equivalence of Wald and likelihood ratio test statistics to replace $\bar{w}$ with $\bar{l}$ and $\bar{W}$ with $\bar{L}$ in (34) and (35). The resulting test statistic is $\hat{S}_f$ as defined in (12). Since $F_{k,w(r_f)}$ is the reference distribution for the Wald statistic, we use $F_{k,w(\hat{r}_f)}$ as the reference distribution for the asymptotically equivalent $\hat{S}_f$.

3.2. Tests for partially synthetic data

The derivations for partially synthetic data follow similarly and hence are presented with less detail. Extending the results in Reiter (2003) for scalar Q to multi-component Q,

$$(Q \mid d^m, B) \sim \text{N}(\bar{q}_m, B/m + \bar{v}_m), \tag{36}$$

$$((m-1) b_m B^{-1} \mid d^m) \sim \text{Wishart}_{m-1}, \tag{37}$$

where $B = \text{Var}(q_i \mid D)$. The data need not be multivariate normal; only the posterior distribution of Q need be. Posterior inferences for $(Q \mid d^m)$ are obtained by integrating the product of (36) and (37) over B.

Assuming that $B = R_p \bar{v}_m$, where R_p is a scalar, the closed-form approximation for the p-value for the Wald test of $Q = Q_0$ is

$$P_p = \int \Pr(\chi_k^2 > (\bar{q}_m - Q_0)^t \bar{v}_m^{-1} (\bar{q}_m - Q_0)(R_p/m + 1)^{-1}) f(R_p \mid d^m) \, dR_p. \tag{38}$$

Using (37) and the assumption $B = R_p \bar{v}_m$,

$$(m r_p t / R_p \mid d^m) \sim \chi_t^2, \tag{39}$$

where r_p is as defined in (7), so that (38) becomes

$$P_p = \Pr(\chi_k^2 > (\bar{q}_m - Q_0)^t \bar{v}_m^{-1} (\bar{q}_m - Q_0)(r_p t/\chi_t^2 + 1)^{-1}) \tag{40}$$

$$= \Pr((\chi_k^2/k)[(r_p t/\chi_t^2 + 1)/(r_p + 1)] > S_p). \tag{41}$$

The reference distribution is obtained as in the fully synthetic data Wald test by matching the first two moments of $(\chi_k^2/k)[(r_p t/\chi_t^2 + 1)/(r_p + 1)]$ to a multiple of an $F_{k,w}$ distribution. The resulting value for w is given in (9). The likelihood ratio test can be derived using the methods outlined in Section 3.1.

4. Simulation studies

This section illustrates the performance of these significance tests using simulation studies. Section 4.1 describes a study in which the imputer generates fully synthetic data, and the analyst uses S_f as the test statistic. Section 4.2 describes a study in which the imputer generates partially synthetic data for all values of one survey variable, leaving the others at their observed values, and the analyst uses $\hat{S}_p$ as the test statistic. For illustrations, the simulations use artificial data and correct posterior distributions for imputations. Of course, in real settings the correct imputation model typically is not known and must be estimated using the observed data and subject-matter expertise. For all simulations, the population sizes are considered infinite so that finite population correction factors are ignored.

4.1. Fully synthetic data

Each observed data set, D, comprises $n = 1000$ multivariate observations drawn randomly from $X \sim N(0, \Sigma_k)$, where $X = (X_0, X_1, \ldots, X_k)$ and Σ_k equals a $(k+1)\times(k+1)$ identity matrix. The simulation uses three dimension scenarios, $k \in (2, 10, 20)$, and three imputation scenarios, $m \in (5, 10, 20)$. All d_i comprise 1000 fully synthetic units, with values drawn from the standard Bayesian posterior predictive distribution, $f(X \mid D)$, using flat priors (see Gelman et al., 1995, Chapter 3).

For each D and attached synthetic data d^m, we fit the regression of X_0 on $(X_1, \ldots, X_k)$. We then test the null hypotheses that all k predictors in the regression have coefficients equal to zero, using the p-values for the significance test based on S_f and $F_{k,w(r_f)}$ as described in Section 2. The simulation is repeated independently 10,000 times for each of the nine combinations of k and m. Table 1 displays the percentages of simulated p-values less than $\alpha = 0.10$, 0.05, and 0.01. The procedures are well calibrated, with most simulated levels within 1% of their corresponding α levels. The procedure is least effective when $k = 2$ and $m = 5$. This suggests it may be possible to improve the procedures by developing reference distributions that do not rely on the normal approximations used in (19) and (20). This is an area for further research.

Table 1
Simulated significance levels for fully synthetic data

Scenario	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$m = 5$			
$k = 2$	0.118	0.072	0.025
$k = 10$	0.098	0.044	0.006
$k = 20$	0.098	0.047	0.006
$m = 10$			
$k = 2$	0.115	0.065	0.019
$k = 10$	0.097	0.045	0.007
$k = 20$	0.102	0.051	0.009
$m = 20$			
$k = 2$	0.108	0.056	0.010
$k = 10$	0.102	0.048	0.010
$k = 20$	0.101	0.055	0.010

Each simulated level is based on 10,000 independent simulation runs.

4.2. Partially synthetic data

Each observed data set, D, comprises $n = 1000$ values of three categorical variables, (Y_1, Y_2, Y_3), generated from the following distributions:

$$\Pr(Y_1 = y) = 0.25, \quad y \in \{1, 2, 3, 4\}, \tag{42}$$

$$\Pr(Y_2 = 1) = e^{-1+0.5Y_1}/(1 + e^{-1+0.5Y_1}), \quad \Pr(Y_2 = 0) = 1 - \Pr(Y_2 = 1), \tag{43}$$

$$\Pr(Y_3 = 1) = e^{g(Y_1, Y_2)}/(1 + e^{g(Y_1, Y_2)}), \quad \Pr(Y_3 = 0) = 1 - \Pr(Y_3 = 1), \tag{44}$$

where $g(Y_1, Y_2) = I(Y_1 = 1) + 0.5I(Y_1 = 2) - 0.5I(Y_1 = 3) - I(Y_1 = 4) - 0.5I(Y_2 = 0) + 0.5I(Y_2 = 1)$, and the notation $I(\ldots)$ represents an indicator variable that equals one when the expression in $(\ldots)$ is true and equals zero otherwise. We plan to test the null hypothesis that the coefficients of the interactions between the levels of Y_1 and Y_2 equal zero, which is true in this simulation design, and to test the null hypothesis that the coefficients of Y_1 equal zero, which is not true in this simulation design.

We generate partially synthetic data to replace all of Y_3, leaving Y_1 and Y_2 at their observed values. To generate synthetic data, we use the posterior predictive distribution $f(Y_3 \mid Y_1, Y_2)$ which is determined as follows. Let ij index the units with $Y_1 = i$ and $Y_2 = j$; let n_{ij} be the number of units in category ij; and, let c_{ij} be the number of units in category ij with $Y_3 = 1$. For all ij, assume $(c_{ij} \mid \pi_{ij}) \sim \text{Bin}(n_{ij}, \pi_{ij})$, and $\pi_{ij} \sim \text{Beta}(1, 1)$. To draw Y_3,

1. For all ij, draw π_{ij} from its posterior distribution, $\pi_{ij} \sim \text{Beta}(c_{ij} + 1, n_{ij} - c_{ij} + 1)$.
2. Draw Y_3 for each unit in category ij from Bernoulli(π_{ij}).

Table 2
Simulated significance levels for saturated versus independence models

Scenario	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$m = 5$	0.081	0.038	0.007
$m = 10$	0.085	0.040	0.008
$m = 20$	0.079	0.039	0.006

Each simulated level is based on 10,000 independent simulation runs.

The simulation includes three imputation scenarios, $m \in \{5, 10, 20\}$, and 10,000 independent D are simulated for each scenario.

For each D and associated d^m, we fit three logistic regressions using Y_3 as the outcome and functions of Y_1 and Y_2 as predictors. The saturated model includes all main effects and interactions of Y_1 and Y_2. The independence model includes only the main effects of Y_1 and Y_2. The third model includes only the main effect of Y_2. Using likelihood ratio tests, we compare the saturated model to the independence model, and the independence model to the model with Y_2 only. Based on (42)–(44), the independence model should fit the data well, and the model with Y_2 only should be rejected in favor of the independence model. The likelihood ratio tests are performed using the test statistic in (13), with the reference distribution $F_{3,w(r_p)}$.

All synthetic and observed data p-values for the test of the independence model versus the model with Y_2 only are extremely small (on the order of 10^{-6}), correctly indicating that the model with Y_2 only does not fit the data well relative to the independence model. For the test of the saturated versus independence models, the percentages of simulated p-values less than $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.01$ are displayed in Table 2. Once again, the procedures are reasonably well calibrated, with most simulated levels within 2% of their corresponding α levels. The departures from the nominal α levels arise primarily because the assumption of $B_m = R_p \bar{v}_m$ is only approximately true.

5. Concluding remarks

In the simulation of Section 4.2, the likelihood ratio test had nearly 100% power for rejecting the Y_2 only model in favor of the correct model. Of course, power depends on many characteristics of the data setting, and extremely high power cannot be expected to exist generally. To get a sense of the power properties of the tests presented here, we can turn to the results of Li et al. (1991), who examined the power properties of large sample significance tests for multiple imputation. Their tests are very similar to those presented here, and they are derived from similar assumptions and approximations. Based on extensive simulation studies, Li et al. (1991) report that power curves for their tests are similar to the power curves for Wald tests based on the observed data. The greatest losses in power occur when the data deviate substantially from the proportionality assumption, which recall in our setting is $B_m = R_f \bar{v}_m$ or $B_m = R_p \bar{v}_m$. The losses are largest when m is small, and mostly disappear for large m.

To conclude, this paper adds significance testing for multi-component estimands to the inferential methods available for synthetic data, thereby increasing the utility of synthetic data approaches. Such approaches offer great promise to guard confidentiality and provide acceptable data utility. Indeed, given current trends in restriction of public access to microdata, it is not inconceivable that simulated data sets may become the only form of releasable, usable microdata. Of course, the key to the success of synthetic data approaches is good-fitting imputation models, and research on specification of such models would complement recent theoretical advances.

Acknowledgements

This work was supported by the United States Bureau of the Census through a contract with Datametrics Research. The author thanks Trivellore Raghunathan, Donald Rubin, and Laura Zayatz for providing statistical support and general motivation for this research, and the referees and editors for their valuable comments and suggestions.

References

Abowd, J.M., Woodcock, S.D., 2001. Disclosure limitation in longitudinal linked data. In: Doyle, P., Lane, J., Zayatz, L., Theeuwes, J. (Eds.), Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. North-Holland, Amsterdam, pp. 215–277.

Dandekar, R.A., Cohen, M., Kirkendall, N., 2002a. Sensitive micro data protection using Latin hypercube sampling technique. In: Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases. Springer, Berlin, pp. 117–125.

Dandekar, R.A., Domingo-Ferrer, J., Sebe, F., 2002b. LHS-based hybrid microdata versus rank swapping and microaggregation for numeric microdata protection. In: Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases. Springer, Berlin, pp. 153–162.

Fienberg, S.E., Steele, R.J., Markov, U.E., 1996. Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: data swapping and log-linear models. In: Proceedings of Bureau of Census 1996 Annual Research Conference, pp. 87–105.

Fienberg, S.E., Markov, U.E., Steele, R.J., 1998. Disclosure limitation using perturbation and related methods for categorical data. J. Official Statist. 14, 485–502.

Franconi, L., Stander, J., 2002. A model based method for disclosure limitation of business microdata. Statistician 51, 1–11.

Franconi, L., Stander, J., 2003. Spatial and non-spatial model-based protection procedures for the release of business microdata. Statist. Comput. 13, 295–306.

Fuller, W.A., 1993. Masking procedures for microdata disclosure limitation. J. Official Statist. 9, 393–406.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman & Hall, London.

Kennickell, A.B., 1997. Multiple imputation and disclosure protection: the case of the 1995 Survey of Consumer Finances. In: Alvey, W., Jamerson, B. (Eds.), Record Linkage Techniques, 1997. National Academy Press, Washington, DC, pp. 248–267.

Li, K.H., 1985. Hypothesis testing with multiple imputation—with emphasis on mixed-up frequencies in contingency tables. Ph.D. Thesis, Department of Statistics, University of Chicago.

Li, K.H., Raghunathan, T.E., Rubin, D.B., 1991. Large sample significance levels from multiply imputed data using moment-based statistics and an f reference distribution. J. Amer. Statist. Assoc. 86, 1065–1073.

Little, R.J.A., 1993. Statistical analysis of masked data. J. Official Statist. 9, 407–426.

Liu, F., Little, R.J.A., 2002. Selective multiple imputation of keys for statistical disclosure control in microdata. In: ASA Proceedings of the Joint Statistical Meetings, pp. 2133–2138.

Meng, X.I., Rubin, D.B., 1992. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika 79, 103–111.

Polettini, S., 2003. Maximum entropy simulation for microdata protection. Statist. Comput. 13, 307–320.

Polettini, S., Franconi, L., Stander, J., 2002. Model-based disclosure protection. In: Domingo-Ferrer, J. (Ed.), Inference Control in Statistical Databases. Springer, Berlin, pp. 83–96.

Raghunathan, T.E., 1987. Large sample significance levels from multiply-imputed data. Ph.D. Thesis, Department of Statistics, Harvard University.

Raghunathan, T.E., Reiter, J.P., Rubin, D.B., 2003. Multiple imputation for statistical disclosure limitation. J. Official Statist. 19, 1–16.

Reiter, J.P., 2002. Satisfying disclosure restrictions with synthetic data sets. J. Official Statist. 18, 531–544.

Reiter, J.P., 2003. Inference for partially synthetic, public use microdata sets. Surv. Methodology 29, 181–188.

Reiter, J.P., 2004. Releasing multiply-imputed, synthetic public use microdata: an illustration and empirical study. J. Roy. Statist. Soc. Ser. A, forthcoming.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Rubin, D.B., 1993. Discussion: statistical disclosure limitation. J. Official Statist. 9, 462–468.

Schafer, J.L., 2000. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Willenborg, L., de Waal, T., 2001. Elements of Statistical Disclosure Control. Springer, New York.