Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Multiple imputation with census data

by Satkartar K. Kinney

December 2012

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

This product is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2012
Vol. 38, No. 2, pp. 215-218
Statistics Canada, Catalogue No. 12-001-X

Multiple imputation with census data

Satkartar K. Kinney [1]

Abstract

A benefit of multiple imputation is that it allows users to make valid inferences using standard methods with simple combining rules. Existing combining rules for multivariate hypothesis tests fail when the sampling error is zero. This paper proposes modified tests for use with finite population analyses of multiply imputed census data for the applications of disclosure limitation and missing data and evaluates their frequentist properties through simulation.

Key Words: Finite Populations; Missing data; Significance testing; Synthetic data.

1. Introduction

Multiple imputation was first proposed for handling nonresponse in large complex surveys (Rubin 1987). Several other uses for multiple imputation have since been proposed, including statistical disclosure limitation and measurement error. An appeal of multiple imputation is that standard methods can be applied to each imputed dataset and then simple combining rules applied, which vary between applications. See Reiter and Raghunathan (2007) for a detailed overview of the different rules and applications. Existing multiple imputation combining rules were developed for use with random samples and superpopulation models (Deming and Stephan 1941). In finite population analyses of census data, where the sampling variance is zero, the combining rules for univariate estimands can still be applied as a special case; however, hypothesis tests for multivariate estimands break down.

Motivated by the use of multiple imputation to generate partially synthetic data (Rubin 1993; Little 1993) for the U.S. Census Bureau's Longitudinal Business Database (Kinney, Reiter, Reznek, Miranda, Jarmin and Abowd 2011), an economic census, this paper derives a multivariate test for finite populations for use with partially synthetic data and extends it to the application of missing data. Extensions to other multiple imputation applications are expected to be straightforward.

The remainder of this paper is organized as follows. Section 2 describes the case of partially synthetic data and Section 3 presents the extension to missing data. Simulations in Section 4 evaluate the combining rules for both the missing data and partially synthetic data cases.

2. Partially synthetic data

Partially synthetic datasets are constructed by replacing selected values in the confidential data with m independent draws from their posterior predictive distribution. For a finite population of size N, let $Z_j = 1$, $j = 1, ..., N$ indicate that unit j has been selected to have any observed values replaced with imputations. Imputations should only be made from the posterior predictive distribution of those units with $Z_j = 1$. For simplicity, in this paper, we assume $Z_j = 1$, $j = 1, ..., N$. Let $Y = (y_1, ..., y_d)$ be the matrix of confidential variables that will be replaced with imputations and X the matrix of variables that will not be replaced. Let $D_{\text{cen}} = (X, Y)$ represent a census of all N units containing confidential data and assume that all units are fully observed, *i.e.*, no missing values are present. Let $Y_{\text{rep}}^{(i)}, i = 1, ..., m$ be the i^{th} imputation of Y, and let $D_{\text{syn}}^{(i)} = (X, Y_{\text{rep}}^{(i)})$. The set $D_{\text{syn}} = \{D_{\text{syn}}^{(i)}, i = 1, ..., m\}$ is what is released to the public.

Any proper imputation procedure from the broad literature on multiple imputation may be used to generate D_{syn} from D_{cen}. The finite population methods proposed here can be used regardless of whether a finite population was assumed in the generation of D_{syn}. Under a finite population assumption, since the data are a fully observed census the imputation model parameters would be considered known and fixed. See Reiter and Kinney (2012) for an illustration of how valid inferences are obtained from partially synthetic random samples generated with both fixed and random imputation model parameters. Simulations (not shown) confirm the same is true in the finite population case.

An analyst with access to D_{syn} but not D_{cen} can obtain valid inferences for a scalar or vector estimand Q using the following quantities:

$$\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m} Q^{(i)} \tag{2.1}$$

$$\bar{U}_m = \frac{1}{m}\sum_{i=1}^{m} U^{(i)} \tag{2.2}$$

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m} (Q^{(i)} - \bar{Q}_m)\,(Q^{(i)} - \bar{Q}_m)' \tag{2.3}$$

1. Satkartar K. Kinney, National Institute of Statistical Sciences, Research Triangle Park, NC 27709, U.S.A. E-mail: saki@niss.org.

where $Q^{(i)}$, $i = 1, ..., m$, is the point estimate of Q obtained from $D_{\text{syn}}^{(i)}$, $U^{(i)}$ is the estimated variance of Q, and B_m is the sample variance of the $Q^{(i)}$, $i = 1, ..., m$.

When there is no sampling variance the combining rules for scalar Q derived by Reiter (2003) can be applied as a special case where $\bar{U}_m = 0$. The resulting simplification means the approximations of Reiter (2003) are not needed and the exact posterior under multivariate normal theory is $(Q \mid D_{\text{syn}}) \sim t_{m-1}(\bar{Q}_m, B_m / m)$. For a vector Q, however, the hypothesis test of Reiter (2005) relies on the assumption that B_∞ is proportional to $\bar{U}_\infty$, *i.e.*, the proportion of information replaced with imputations is the same across components of Q, so a different assumption is needed for the case $\bar{U}_\infty = 0$.

2.1 Proposed multivariate test

In this section an alternate test is derived based on the stronger assumption that $B_\infty = r_\infty I$, for a scalar quantity r_∞ and k-dimensional identity matrix I. In other words, the between-imputation variance is constant across components of Q, and B_∞ is assumed to be diagonal. In both the Reiter (2005) test and the proposed test, one averages across variance components so the test is moderately robust to this assumption; however, the randomization validity declines when the estimates of Q, $\bar{Q}^{(i)}$, $i = 1, ..., m$, are highly correlated. This is evaluated with simulations in Section 4.3. Comparable tests based on the assumption $B_\infty \propto \bar{U}_\infty$ are known to lose power when the assumption is not met (Li *et al.* 1991).

The proposed test for the hypothesis H_0: $Q = Q_0$ is conducted by referring the test statistic

$$S_c = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{k r_c}$$

to an $F_{k,k(m-1)}$ distribution, where $r_c = 1/m \; tr(B_m) / k$.

Under the assumption $B_\infty = r_\infty I$, the Bayesian p-value is given by

$$\int P(\chi_k^2 > (Q_0 - \bar{Q})' T_\infty^{-1} (Q_0 - \bar{Q}) \mid D_{\text{syn}}, B_\infty) \; P(B_\infty \mid D_{\text{syn}}) \, dB_\infty \quad (2.4)$$

$$= \int P\left(\chi_k^2 > \frac{(Q_0 - \bar{Q})' I (Q_0 - \bar{Q})}{r_\infty / m} \mid D_{\text{syn}}, r_\infty \right) P(r_\infty \mid D_{\text{syn}}) \, dr_\infty$$

$$= \int P\left(\frac{\chi_k^2}{k} \cdot \frac{r_\infty}{m r_c} > S_c \mid D_{\text{syn}}, \; r_\infty \right) P(r_\infty \mid D_{\text{syn}}) \, dr_\infty. \quad (2.5)$$

Thus the proportionality assumption reduces the number of variance parameters to be estimated from $k(k-1)/2$ to 1 and allows for the closed-form approximation of the integral in (2.4). As $\bar{U}_\infty = 0$, the derivation is simplified from Reiter (2005). To complete the integration, we need the distribution of $(r_\infty \mid D_{\text{syn}})$. Extending the scalar case in Reiter (2003), the sampling distribution of $Q^{(i)}$, the estimate of Q obtained from $D_{\text{syn}}^{(i)}$, is given by $(Q^{(i)} \mid Q_{\text{cen}}, B_\infty) \sim N(Q_{\text{cen}}, B_\infty)$. Under the proportionality assumption, this becomes $(Q^{(i)} \mid Q_{\text{cen}}, r_\infty) \sim N(Q_{\text{cen}}, r_\infty I)$. With diffuse priors and standard multivariate normal theory for sample covariance matrices, we obtain

$$(m-1) \frac{\sum_{i=1}^{m} (Q^{(i)} - \bar{Q}_m)(Q^{(i)} - \bar{Q}_m)'}{(m-1) r_\infty} \mid D_{\text{syn}} \sim \text{Wish}(m-1, I).$$

Taking the trace of each side and integrating over r_∞ in (2.5) yields a Bayesian p-value of

$$P\left(\frac{\chi_k^2}{k} \frac{k(m-1)}{\chi_{k(m-1)}^2} > S_c \mid D_{\text{syn}} \right) = P(F_{k,k(m-1)} > S_c \mid D_{\text{syn}}).$$

3. Missing data

The extension to missing data is straightforward. When $\bar{U}_\infty = 0$, the combining rules (Rubin 1987) for scalar estimands q simplify so that $(q \mid D_{\text{com}}) \sim N(\bar{q}_m, (1 + 1/m) B_m)$, where D_{com} is the set of m completed datasets. Similar to Section 2, the tests of Rubin (1987) and Li, Raghunathan and Rubin (1991) for multivariate components rely on the assumption that $B_\infty \propto \bar{U}_\infty$, and so when $\bar{U}_\infty = 0$ we derive a test under the assumption $B_\infty = r_\infty I$.

Following derivation procedures similar to that of Section 2.1, the Bayesian p-value for testing H: $Q = Q_0$ with k-dimensional Q is found to be $P(F_{k,k(m-1)} > S_q \mid D_{\text{com}})$ where

$$S_q = \frac{(Q_0 - \bar{Q}_m)'(Q_0 - \bar{Q}_m)}{k r_q},$$

and $r_q = (1 + 1/m) \; tr(B_m) / k$.

4. Simulation study

In this section, simple simulation examples illustrate the analytic validity of the proposed combining rules, first for the case of partially synthetic data, and then for the case missing data. Lastly, the robustness of the tests to the proportionality assumption is evaluated.

For a population of $N = 50{,}000$, $X = (X_1, ..., X_{20})$ is drawn from a multivariate normal distribution with mean zero and covariance matrix with 1 in each diagonal element and 0.5 in each off-diagonal element. Y is drawn from a standard normal distribution. For each of 5,000 iterations, a new finite population is generated and m imputations are drawn for $m \in \{2, 5, 10\}$. The proposed hypothesis tests are conducted for H_0: $Q = Q_0$, where Q is the vector of regression coefficients, excluding the intercept, of the regression of Y on X and has dimension k, $k \in \{2, 5, 20\}$, and Q_0 is the true value of Q determined from the finite population (X, Y). Since H_0 is true by design, H_0 should be rejected $100\alpha\%$ of the time, for significance level $\alpha = 0.05$.

Random sampling scenarios are also simulated for comparison purposes. At each iteration, a random sample of size $s = 50{,}000$ from an infinite population is generated from the distributions described above, prior to generating the m missing data and synthetic imputations. The same hypothesis H_0: $Q = Q_0$ is tested where Q_0 is the vector of true population values. The combining rules for the hypothesis tests are those of Reiter (2005) in the synthetic data case and Li *et al.* (1991) and Rubin (1987) in the missing data case.

4.1 Partially synthetic data imputations

Let Y be a confidential response variable and X be unreplaced predictors. Then Y_{syn} is generated by taking m independent draws from the posterior predictive distribution $f(Y \mid X)$ assuming a normal linear model, using all available data.

Table 1 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to the significance level 0.05, and close to the random sampling results. From these results it appears that the proposed combining rules for population data have good frequentist properties. Not shown are the rejection rates when the rules from random samples (Reiter 2005) were applied to finite populations, which were observed to be quite high, typically 1, in the simulations conducted.

Table 1
Comparison of nominal 5% rejection rates for tests on partially synthetic data

	$k = 2$	$k = 5$	$k = 20$
Census data			
$m = 2$	0.048	0.065	0.052
$m = 5$	0.048	0.061	0.057
$m = 10$	0.051	0.067	0.055
Random sampling			
$m = 2$	0.067	0.062	0.060
$m = 5$	0.054	0.052	0.050
$m = 10$	0.047	0.049	0.049

4.2 Missing data

Simulations analogous to the synthetic data simulations were conducted for the missing data case. The missing values of Y are imputed from the posterior predictive distribution $f(Y_{\text{obs}} \mid X)$ assuming a normal linear model. Missingness is simulated to be completely at random, with $P(R_l = 1) = 0.3$, $l = 1, ..., s$, where R is an indicator variable for missingness.

Table 2 gives the nominal 5% rejection rate for the proposed hypothesis test for multicomponent estimands, which are seen to be close to 0.05, and to the random sampling results. From these results it appears that the proposed combining rules for population data yield valid inferences.

Table 2
Comparison of nominal 5% rejection rates for tests using completed census data

	$k = 2$	$k = 5$	$k = 20$
Census data			
$m = 2$	0.052	0.061	0.053
$m = 5$	0.048	0.063	0.051
$m = 10$	0.048	0.058	0.054
Random sampling			
$m = 2$	0.061	0.056	0.053
$m = 5$	0.056	0.052	0.052
$m = 10$	0.048	0.050	0.051

4.3 Robustness

The assumption that $B_\infty \propto r_\infty I$ is striking at first glance, and is unlikely to be exactly true. In this section we evaluate the effect of strong correlations across components of Q. While moderately strong correlations were present in the previous simulations, here we increase the magnitude of the between-imputation variance, increasing the magnitude of the differences across the diagonal of B as well as the distance from zero of the off-diagonal elements of B.

These simulations are set up as before, for the finite population case, with $k = 5$ and $m = 5$. The population in each iteration is generated in the same way as before, except that we let $Y = (1, 2, 5, 10, 20, 0, ..., 0)(X_1, X_2, ..., X_{20})' + \eta$, $\eta \sim N(0, 100)$ and $X_2 = c \cdot X_1 + \varepsilon$, $c \in \{0.5, 1, 5\}$ and $\varepsilon \sim N(0, 1)$. Increasing values of c yields increasingly higher correlations. The large variance for η induces larger and more variable values for elements of B.

The results in Table 3 indicate that while the tests have good properties even with moderately high violations of the proportionality assumption, their performance declines with increasingly large correlations. Continuing our assumption that Q represents a vector of regression coefficients, presence of such large correlation may also be indicative of multicollinearity in the model at hand, so analysts faced with high correlation across $\overline{Q}^{(i)}$ might take steps to reduce multicollinearity before applying the proposed tests. If

variables are of substantially differing magnitude, standardization to rescale them will reduce differences across Q.

Table 3
Evaluation of tests under assumption violations, $k = 5, m = 5$

	$c = 0.5$	$c = 1$	$c = 5$
Synthetic Data	0.059	0.083	0.145
Missing Data	0.051	0.083	0.136

Acknowledgements

A portion of this work was conducted while the author was a student at Duke University, supported by NSF grant ITR-0427889 and under the guidance of Jerry Reiter, whose assistance is greatly appreciated. In addition, the comments of anonymous reviewers were quite helpful.

References

Deming, W.E., and Stephan, F.F. (1941). On the interpretation of censuses as samples. *Journal of the American Statistical Association*, 36, 213, 45-49.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S. and Abowd, J.M. (2011). Toward unrestricted public-use business microdata: The Longitudinal Business Database. *International Statistical Review*, 79, 3, 362-384.

Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991). Large-sample significance levels from multiply-imputed data using moment-based statistics and an *F* reference distribution. *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 2, 181-188.

Reiter, J.P. (2005). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365-377.

Reiter, J.P., and Kinney, S.K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. Technical report, National Institute of Statistical Sciences.

Reiter, J.P., and Raghunathan, T.E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102, 1462-1471.