Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 2283–2302
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Confidence intervals for marginal parameters under imputation for item nonresponse☆

Yongsong Qin[a], J.N.K. Rao[b,*], Qunshu Ren[c]

[a]*School of Mathematical Sciences, Guangxi Normal University, Guilin, Guangxi 541004, China*
[b]*School of Mathematics and Statistics, Carleton University, Ottawa, Ont., Canada K1S 5B6*
[c]*Canada Institute for Health Information, Ottawa, Ont., Canada K2A 4H6*

Received 24 July 2006; received in revised form 5 June 2007; accepted 11 October 2007
Available online 17 November 2007

Abstract

Item nonresponse occurs frequently in sample surveys and other approaches to data collection. We consider three different methods of imputation to fill in the missing values in a random sample $\{Y_i,\ i = 1, \ldots, n\}$: (i) mean imputation ($M$), (ii) random hot deck imputation (R), and (iii) adjusted random hot deck imputation (A). Asymptotic normality of the imputed estimators of the mean μ under M, R and A and the distribution function $\theta = F(y)$ and qth quantile θ_q, under R and A is established, assuming that the values are missing completely at random. This result is used to obtain normal approximation (NA)-based confidence intervals on μ, θ and θ_q. In the case of θ_q, Woodruff [1952. Confidence intervals for medians and other position measures, J. Amer. Statist. Assoc. 47, 635–646]-type confidence intervals are also obtained under R and A. Empirical log-likelihood ratios for the three cases are also obtained and shown to be asymptotically scaled χ^2_1. This result is used to obtain asymptotically correct empirical likelihood (EL)-based confidence intervals on μ, θ and θ_q. Results of a simulation study on the finite sample performance of NA-based and EL-based confidence intervals are reported. Confidence intervals obtained here do not require identification flags on the imputed values in the data file; only the estimated response rate is needed with the imputed data file. This feature of our method is important because identification flags often may not be provided in practice with the data file due to confidentiality reasons.

MSC: primary 62G05; secondary 62E20

Keywords: Distribution function; Empirical likelihood; Mean; Quantile; Random imputation

1. Introduction

Item nonresponse occurs frequently in sample surveys and other approaches to data collection. Reasons for item nonresponse include unwillingness of sampled units to respond on some items, failure of the investigator to gather correct information on certain items, loss of item values caused by uncontrollable factors, and so on. Item nonresponse is usually handled by some form of imputation to fill in missing item values. Brick and Kalton (1996) list the main advantages of imputation over other methods for handling missing data. Imputation permits the creation of a general-

☆ Supported by a grant from the Natural Sciences and Engineering Research Council of Canada and the National Natural Science Foundation of China (10661003).
* Corresponding author. Tel.: +1 613 829 6555; fax: +1 613 520 3536.
E-mail address: jrao@math.carleton.ca (J.N.K. Rao).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.10.009

purpose complete public-use data file with or without identification flags on the imputed values that can be used for standard analyses, such as the calculation of item means (or totals), distribution functions and quantiles. Secondly, analyses based on the imputed data file are internally consistent. Thirdly, imputation retains all the reported data in multivariate analyses.

In this paper, we focus on marginal imputation for each item in the case of simple random sampling; extension to stratified random sampling with independent imputations across strata is also outlined. We study commonly used mean imputation and random (hot deck) imputation of donor values for each item. We also study a new method, called adjusted random imputation (Chen et al., 2000). We assume missing completely at random (MCAR) mechanism for each item. Random imputation preserves the distribution of item values and the resulting imputed estimators of mean, distribution function and quantile are asymptotically consistent, but it leads to imputation variance which can be a significant component of the total variance of the imputed estimators if the item response rate is not high. Mean imputation eliminates the imputation variance in the case of mean, but the distribution of item values is not preserved because of the spike at the common imputed value. As a result, the imputed estimators of distribution function and quantile are inconsistent. Adjusted random imputation eliminates the imputation variance and at the same time preserves the distribution of item values, leading to consistent estimators of distribution functions and quantiles.

Analysts often treat the imputed values as actual values and calculate the estimates, standard errors and confidence intervals. But this can lead to significant underestimation of variance and confidence interval undercoverage due to ignoring the variability associated with the imputed values. In this paper, we develop asymptotically valid inferences that take account of imputation. In particular, we establish the asymptotic normality of the imputed estimators and construct normal approximation (NA)-based confidence intervals on item mean, distribution function and quantile. We also obtain empirical likelihood (EL)-based confidence intervals. In the complete data setting, the original idea of EL dates back to Hartley and Rao (1968) in the context of sample surveys, and Owen (1988, 1990) made a systematic study of the EL method. EL confidence intervals are range preserving and transformation respecting and the shape and orientation of EL intervals are determined entirely by the data, unlike the NA-based intervals. However, the EL method requires modifications in the case of data with imputed values.

We assume simple random sampling from a large population of size N and negligible sampling fraction n/N. We also focus on a single item Y and associated mean $\mu = E(Y)$, distribution function $\theta = F(y) = P(Y \leqslant y)$ for given $y \in R$ and qth quantile $\theta_q = F^{-1}(q)$, $0 < q < 1$. No parametric structure on the distribution of Y is assumed except that $0 < \mathrm{var}(Y) = \sigma^2 < \infty$. The sample of incomplete data $\{(Y_i, \delta_i);\ i = 1, 2, \ldots, n\}$ may be regarded as an i.i.d. sample generated from the random vector (Y, δ), where $\delta_i = 0$ if Y_i is missing and $\delta_i = 1$ otherwise. We assume that Y is MCAR, i.e., $P(\delta = 1|Y) = P(\delta = 1) = p$, $0 < p \leqslant 1$. In the stratified case, MCAR is assumed within strata but the probability of response can vary across strata.

Let $r = \sum_{i=1}^{n} \delta_i$ and $m = n - r$. Denote the set of respondents as s_r, the set of nonrespondents as s_m, $s = s_r \cup s_m$, and the mean of respondents as $\bar{Y}_r = \frac{1}{r}\sum_{i \in s_r} Y_i$. We consider three imputation methods: mean imputation (M), random hot deck imputation (R) and adjusted random hot deck imputation (A). Let $Y_i^{(M)}$, $Y_i^{(R)}$ and $Y_i^{(A)}$, $i \in s_m$, be the imputed values for the missing data based on M, R and A, respectively. Mean imputation uses $\bar{Y}_r$ as the imputed value, i.e., $Y_i^{(M)} = \bar{Y}_r$ for all $i \in s_m$. Random hot deck imputation selects a simple random sample of size m with replacement from s_r and then uses the associated Y-values as donors, that is, $Y_i^{(R)} = Y_j$ for some $j \in s_r$. The adjusted random imputation method, proposed by Chen et al. (2000), uses $Y_i^{(A)} = \bar{Y}_r + (Y_i^{(R)} - \bar{Y}_m^{(R)})$ as imputed values, where $\bar{Y}_m^{(R)} = \frac{1}{m}\sum_{i \in s_m} Y_i^{(R)}$. Let

$$Y_{M,i} = \delta_i Y_i + (1 - \delta_i) Y_i^{(M)}, \quad Y_{R,i} = \delta_i Y_i + (1 - \delta_i) Y_i^{(R)}, \quad Y_{A,i} = \delta_i Y_i + (1 - \delta_i) Y_i^{(A)},$$

$i = 1, \ldots, n$, represent 'completed' data based on M, R and A, respectively.

In Section 2, we establish the asymptotic normality of the imputed estimators under simple random sampling and construct NA-based confidence intervals for the population parameters. Woodruff (1952)-type confidence intervals for the quantiles under R and A are also obtained. In Section 3, EL ratio statistics are constructed, limiting distributions of these statistics are derived, and EL-based confidence intervals for the population parameters are obtained. We show that all the confidence intervals have asymptotically correct coverage accuracy. Results of a simulation study on the relative performance of NA-based and EL-based confidence intervals are reported in Section 4, as well as Woodruff-based intervals for the median, $\theta_{1/2}$. Using the results in Sections 2 and 3 for simple random sampling, extensions to stratified random sampling are outlined in Section 5. Proofs of theorems and lemmas are delegated to an Appendix A.