Confidence Intervals for Marginal Parameters Under Imputation for Item Nonresponse[1]

By Yongsong Qin   J. N. K. Rao[2]   and   Qunshu Ren

Guangxi Normal University and *Carleton University*

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 q-th 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 based confidence intervals on μ, θ and θ_q. In the case of θ_q, Bahadur representations and Woodruff (1952)-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 normal approximation 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.

AMS classifications: Primary 62G05; secondary 62E20

Keywords: Distribution function, empirical likelihood, mean, quantile, random imputation.

[1] Supported by a grant from the Natural Sciences and Engineering Research Council of Canada

[2] Corresponding author.

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-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, Rao and Sitter, 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 estimators if the item response rate is not high. Mean imputation eliminates the imputation variance, 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 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 empirical likelihood dates back to Hartley and Rao (1968) in the context of sample surveys, and Owen (1988, 1990) made a systematic study of the empirical likelihood (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 normal approximation 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 $\theta = F(y) = P(Y \leq y)$ for given $y \in R$ and q-th quantile $\theta_q = F^{-1}(q)$, $0 < q < 1$. No parametric structure on the distribution of Y is assumed except that $0 < \text{var}(Y) = \sigma^2 < \infty$. 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 missing completely at random (MCAR), i.e., $P(\delta = 1|Y) = P(\delta = 1) = p$, $0 < p \leq 1$. 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, 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, Rao and Sitter (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

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

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

In Section 2, we establish the asymptotic normality of the imputed estimators and construct normal-approximation based confidence intervals for the population parameters. Bahadur representationsof quantiles under R and A are also given as well as Woodruff (1952) type confidence intervals for the quantiles. In Section 3, empirical likelihood ratio statistics are constructed, limiting distributions of these statistics are derived, and empirical likelihood 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 normal approximation based and EL based confidence intervals are reported in Section 4, as well as Woodruff-based intervals for the median $\theta_{\frac{1}{2}}$. Extension to stratified random sampling is outlined in Section 5. Proofs are delegated to an Appendix (Section 7).

2. NORMAL APPROXIMATION

2.1 Mean μ

Estimators for μ after imputation under M, R and A are given by

$$\bar{Y}_M = \frac{1}{n}\sum_{i=1}^{n} Y_{M_i}, \quad \bar{Y}_R = \frac{1}{n}\sum_{i=1}^{n} Y_{R,i}, \quad \bar{Y}_A = \frac{1}{n}\sum_{i=1}^{n} Y_{A,i}.$$

It is clear that $\bar{Y}_M = \bar{Y}_A = \bar{Y}_r$.

The result on asymptotic normality of the above estimators for μ is summarized in Theorem 2.1. The proof of Theorem 2.1 is given in the Appendix.

Theorem 2.1 *Assume that* $0 < p \le 1$ *and* $0 < Var(Y) = \sigma^2 < \infty$. *Then*

$$\sqrt{n}(\bar{Y}_M - \mu) \xrightarrow{d} N(0, p^{-1}\sigma^2), \tag{2.1}$$

and

$$\sqrt{n}(\bar{Y}_A - \mu) \xrightarrow{d} N(0, p^{-1}\sigma^2), \tag{2.2}$$

as $n \to \infty$. *Further, assume that there exists an* $\alpha_0 > 0$ *such that* $E|Y|^{2+\alpha_0} < \infty$. *Then, as* $n \to \infty$,

$$\sqrt{n}(\bar{Y}_R - \mu) \xrightarrow{d} N(0, (1 - p + p^{-1})\sigma^2). \tag{2.3}$$

From Theorem 2.1, $\bar{Y}_M, \bar{Y}_R$ and $\bar{Y}_A$ are all consistent estimators of μ. Also, it follows from (2.1), (2.2) and (2.3) that the asymptotic variances of $\bar{Y}_M$ and $\bar{Y}_A$ are equal and smaller or equal to the asymptotic variance of $\bar{Y}_R$. Thus, $\bar{Y}_M$ and $\bar{Y}_A$ have higher asymptotic efficiency (AE) than $\bar{Y}_R$.

To obtain consistent estimators of σ^2 under different imputations, we examine the sample variances of the completed data. Under mean imputation, the sample variance is

$$
\begin{aligned}
s_M^2 &= \frac{1}{n-1}\sum_{i\in s}(Y_{M,i}-\bar{Y}_M)^2 \\
&= \frac{1}{n-1}\sum_{i\in s_r}(Y_i-\bar{Y}_r)^2 = p\sigma^2 + o_p(1).
\end{aligned}
$$

It follows that under mean imputation $\hat{p}^{-1}s_M^2$ is a consistent estimator of σ^2. Secondly, under random imputation, the sample variance is

$$s_R^2 = \frac{1}{n-1}\sum_{i\in s}(Y_{R,i}-\bar{Y}_R)^2.$$

From the proof of Theorem 3.1, we have

$$s_R^2 = \sigma^2 + o_p(1).$$

It follows that under random imputation s_R^2 is a consistent estimator of σ^2. Finally, under adjusted random imputation, the sample variance is

$$s_A^2 = \frac{1}{n-1}\sum_{i\in s}(Y_{A,i}-\bar{Y}_A)^2.$$

From the proof of Theorem 3.1, we have

$$s_A^2 = \sigma^2 + o_p(1).$$

It follows that under adjusted random imputation s_A^2 is a consistent estimator of σ^2. Using Theorem 2.1 and the above estimators of σ^2, we obtain normal approximation based confidence intervals for μ. We assume that the observed response rate $\hat{p} = r/n = \sum_{i=1}^n \delta_i/n$ is reported in the data file. However, we do not need to know which sampled units have imputed values (i.e. individual identification flags, δ_i, are not needed) in the construction of confidence intervals throughout this paper. It is often the case with survey data that identification flags are not provided for confidentiality reasons, among others. Throughout this paper, we take the observed response rate $\hat{p}$ as the estimator of p. It is a consistent estimator of p. Let $X \sim N(0,1)$ and $z_{\alpha/2}$ be such that $P(|X| \leq z_{\alpha/2}) = 1-\alpha$, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of $N(0,1)$. We then have

(1). CI under mean imputation:

$$[\bar{Y}_M - z_{\alpha/2} n^{-1/2} \hat{p}^{-1/2} s_M,\ \bar{Y}_M + z_{\alpha/2} n^{-1/2} \hat{p}^{-1/2} s_M z_{\alpha/2}].$$

(2). CI under random imputation:

$$[\bar{Y}_R - z_{\alpha/2} n^{-1/2} (1 - \hat{p} + \hat{p}^{-1})^{1/2} s_R,\quad \bar{Y}_R + z_{\alpha/2} n^{-1/2} (1 - \hat{p} + \hat{p}^{-1})^{1/2} s_R],$$

and

(3). CI under adjusted random imputation:

$$[\bar{Y}_A - z_{\alpha/2} n^{-1/2} \hat{p}^{-1/2} s_A,\quad \bar{Y}_A + z_{\alpha/2} n^{-1/2} \hat{p}^{-1/2} s_A].$$

The above confidence intervals are asymptotically correct $(1 - \alpha)$-level intervals for the mean μ.

2.2 Distribution Function θ

We only consider random imputation and adjusted random imputation in estimating θ because the usual estimator of θ under mean imputation is not consistent. The estimators of $\theta = F(y)$ under random imputation and adjusted random imputation are respectively given by

$$F_R(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_{R,i} \leq y), \tag{2.4}$$

and

$$F_A(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_{A,i} \leq y). \tag{2.5}$$

The result on the asymptotic normality associated with (2.4) and (2.5) is summarized in Theorem 2.2. The proof of Theorem 2.2 is given in the Appendix.

THEOREM 2.2 *Assume that $F(y) > 0$. Then,*

$$\sqrt{n}(F_R(y) - \theta) \xrightarrow{d} N[0, (1 - p + p^{-1})F(y)\{1 - F(y)\}], \tag{2.6}$$

as $n \to \infty$. Further, assume that there exists an $\alpha_0 > 0$ such that $E|Y|^{2+\alpha_0} < \infty$, and that the density function $f(\cdot)$ of Y exists and continuous in a neighborhood of y. Then,

$$\sqrt{n}(F_A(y) - \theta) \xrightarrow{d} N[0, \sigma^2_{A,F}(y)], \tag{2.7}$$

as $n \to \infty$, where $\sigma^2_{A,F}(y) = (1 + p^{-1} - p)F(y)\{1 - F(y)\} + (1 - p)[f^2(y)\sigma^2 + 2f(y)E\{YI(Y \leq y)\} - 2f(y)F(y)\mu]$.

It follows from Theorem 2.2 that both $F_R(y)$ and $F_A(y)$ are consistent estimators of $F(y)$. To apply Theorem 2.2 for constructing confidence intervals on θ under A, we need the following result which is proved in the Appendix.

LEMMA 2.1 *Under conditions of Theorem 2.2,*

$$\hat{f}_A(y) \equiv \frac{F_A(y + n^{-1/2}) - F_A(y - n^{-1/2})}{2n^{-1/2}} = f(y) + o_p(1).$$

Using Theorem 2.2 and Lemma 2.1, we obtain normal approximation based confidence intervals on θ under R and A: (1). CI under random imputation:

$$\Big[F_R(y) - z_{\alpha/2}n^{-1/2}(1 - \hat{p} + \hat{p}^{-1})^{1/2}\hat{\sigma}_{R,F}(y),$$
$$F_R(y) + z_{\alpha/2}n^{-1/2}(1 - \hat{p} + \hat{p}^{-1})^{1/2}\hat{\sigma}_{R,F}(y)\Big],$$

where $\hat{\sigma}^2_{R,F}(y) = F_R(y)\{1 - F_R(y)\}$. (2). CI under adjusted random imputation:

$$[F_A(y) - z_{\alpha/2}n^{-1/2}\hat{\sigma}_{A,F}(y),\ F_A(y) + z_{\alpha/2}n^{-1/2}\hat{\sigma}_{A,F}(y)],$$

where

$$\begin{aligned}\hat{\sigma}^2_{A,F}(y) &= (1 + \hat{p}^{-1} - \hat{p})F_A(y)\{1 - F_A(y)\} \\ &+(1 - \hat{p})[\hat{f}^2_A(y)\hat{\sigma}^2 + 2\hat{f}_A(y)\hat{E}\{YI(Y \leq y)\} - 2\hat{f}_A(y)F_A(y)\bar{Y}_A]\end{aligned} \tag{2.8}$$

with $\hat{\sigma}^2 = s_A^2$, and $\hat{E}\{YI(Y \leq y)\} = \frac{1}{n}\sum_{i \in s} Y_{A,i} I(Y_{A,i} \leq y)$.

Similar to the proof of Theorem 3.1, it can be shown that $\hat{E}\{YI(Y \leq y)\} = E\{YI(Y \leq y)\} + o_p(1)$. Combining with Lemma 2.1, it is easy to see that $\hat{\sigma}^2_{A,F}(y)$ is a consistent estimator of $\sigma^2_{A,F}(y)$. The above confidence intervals are asymptotically correct $(1-\alpha)$-level intervals on $\theta = F(y)$. Note that the above confidence intervals do not require the identification of imputed values on the data file.

2.3 q-th Quantile θ_q

We only consider random imputation and adjusted random imputation for estimating θ_q because the estimator of θ_q under mean imputation is not consistent. The estimators of $\theta_q = F^{-1}(q)$ after random imputation and adjusted random imputation are respectively given by

$$\hat{\theta}_q^{(R)} = \inf_u\{F_R(u) \geq q\} = F_R^{-1}(q),$$

and

$$\hat{\theta}_q^{(A)} = \inf_u\{F_A(u) \geq q\} = F_A^{-1}(q),$$

where $F_R(u)$ and $F_A(u)$ are defined in (2.4) and (2.5), respectively.

The result on the asymptotic normality associated with $\hat{\theta}_q^{(R)}$ and $\hat{\theta}_q^{(A)}$ is given in Theorem 2.3. The proof of Theorem 2.3 is given in the Appendix.

Theorem 2.3 *Suppose that there exists an $\alpha_0 > 0$ such that $E|Y|^{2+\alpha_0} < \infty$, and that the density function $f(\cdot)$ of Y exists and continuous in a neighborhood of θ_q with $f(\theta_q) > 0$. Then as $n \to \infty$,*

$$\sqrt{n}(\hat{\theta}_q^{(R)} - \theta_q) \xrightarrow{d} N(0, \sigma^2_{R,q}), \tag{2.9}$$

and

$$\sqrt{n}(\hat{\theta}_q^{(A)} - \theta_q) \xrightarrow{d} N(0, \sigma^2_{A,q}), \tag{2.10}$$

where $\sigma^2_{R,q} = (1 - p + p^{-1})q(1 - q)/f^2(\theta_q), \sigma^2_{A,q} = \sigma^2_{A1}/f^2(\theta_q),$ *and* $\sigma^2_{A1} = (1 + p^{-1} - p)q(1-q) + (1-p)[f^2(\theta_q)\sigma^2 + 2f(\theta_q)E\{YI(Y \leq \theta_q)\} - 2f(\theta_q)q\mu]$. *Further, Bahadur representations of* $\hat{\theta}_q^{(R)}$ *and* $\hat{\theta}_q^{(A)}$ *are given by*

$$\hat{\theta}_q^{(R)} = \theta_q - \frac{F_R(\theta_q) - F(\theta_q)}{f(\theta_q)} + o_p(n^{-1/2}), \tag{2.11}$$

and

$$\hat{\theta}_q^{(A)} = \theta_q - \frac{F_A(\theta_q) - F(\theta_q)}{f(\theta_q)} + o_p(n^{-1/2}). \tag{2.12}$$

To apply Theorem 2.3 for constructing confidence intervals on θ_q, we need the following result which is proved in the Appendix.

Lemma 2.2 *Under the conditions of Theorem 2.3,*

$$\hat{f}_R(\hat{\theta}_q^{(R)}) = \frac{F_R(\hat{\theta}_q^{(R)} + n^{-1/2}) - F_R(\hat{\theta}_q^{(R)} - n^{-1/2})}{2n^{-1/2}} = f(\theta_q) + o_p(1),$$

$$\hat{f}_A(\hat{\theta}_q^{(A)}) = \frac{F_A(\hat{\theta}_q^{(A)} + n^{-1/2}) - F_A(\hat{\theta}_q^{(A)} - n^{-1/2})}{2n^{-1/2}} = f(\theta_q) + o_p(1),$$

and

$$\frac{1}{n}\sum_{i=1}^{n} Y_{A,i} I(Y_{A,i} \leq \hat{\theta}_q^{(A)}) = E\{YI(Y \leq \theta_q)\} + o_p(1).$$

Using Lemma 2.2 and Theorem 2.3, we obtain normal approximation based confidence intervals on θ_q:

(1). CI under random imputation:

$$\left[\hat{\theta}_q^{(R)} - z_{\alpha/2} n^{-1/2} \hat{\sigma}_{R,q}, \hat{\theta}_q^{(R)} + z_{\alpha/2} n^{-1/2} \hat{\sigma}_{R,q}\right],$$

where

$$\hat{\sigma}^2_{R,q} = (1 - \hat{p} + \hat{p}^{-1})q(1 - q)/\hat{f}^2_R(\hat{\theta}_q^{(R)}).$$

From Lemma 2.2, we can see that $\hat{\sigma}^2_{R,q}$ is a consistent estimator of $\sigma^2_{R,q}$.

(2). CI under adjusted random imputation:

$$\Big[\hat{\theta}_q^{(A)} - z_{\alpha/2}n^{-1/2}\hat{\sigma}_{A,q}, \hat{\theta}_q^{(A)} + z_{\alpha/2}n^{-1/2}\hat{\sigma}_{A,q}\Big],$$

where

$$\hat{\sigma}^2_{A,q} = \hat{\sigma}^2_{A1,q}/\hat{f}^2_A(\hat{\theta}_q^{(A)}).$$

with

$$\begin{aligned}\hat{\sigma}^2_{A1,q} = (1+\hat{p}^{-1}-\hat{p})q(1-q) + (1-\hat{p})\Big[&\hat{f}^2_A(\hat{\theta}_q^{(A)})\hat{\sigma}^2 \\ &+ 2\hat{f}_A(\hat{\theta}_q^{(A)})\Big\{\frac{1}{n}\sum_{i\in s_r} Y_i I(Y_i \le \hat{\theta}_q^{(A)}) + \frac{1}{n}\sum_{i\in s_m} Y_i^{(A)} I(Y_i^{(A)} \le \hat{\theta}_q^{(A)})\Big\} \\ &- 2\hat{f}_A(\hat{\theta}_q^{(A)})q\bar{Y}_A\Big].\end{aligned} \tag{2.13}$$

From Lemma 2.2, we can see that $\hat{\sigma}^2_{A,q}$ is a consistent estimator of $\sigma^2_{A,q}$. The above confidence intervals are asymptotically correct $(1-\alpha)$-level intervals on θ_q.

Using the ingenious method of Woodruff (1952), different intervals on θ_q under R and A can be constructed. An advantage of Woodruff intervals under R is that the intervals can be obtained from the estimator of $F(y)$ without estimating the density function $f(\theta_q)$.

(W1). Woodruff-type CI under random imputation:

$$\begin{aligned}&[F_R^{-1}(q - z_{\alpha/2}n^{-1/2}\{q(1-q)(1-\hat{p}+\hat{p}^{-1})\}^{1/2}), \\ &F_R^{-1}(q + z_{\alpha/2}n^{-1/2}\{q(1-q)(1-\hat{p}+\hat{p}^{-1})\}^{1/2})].\end{aligned}$$

We note that $n^{-1}q(1-q)(1-\hat{p}+\hat{p}^{-1})$ is a consistent estimator of the variance of $F_R(\theta_q)$. Denote $s_n = \{q(1-q)(1-\hat{p}+\hat{p}^{-1})\}^{1/2}$. We now show that as $n\to\infty$, the above Woodruff confidence interval is asymptotically correct, i.e.,

$$P\Big[F_R^{-1}(q - z_{\alpha/2}n^{-1/2}s_n) \le \theta_q \le F_R^{-1}(q + z_{\alpha/2}n^{-1/2}s_n)\Big] \to 1-\alpha. \tag{2.14}$$

Similar to the proof of Lemma 2.1 in the Appendix, we can show that

$$F_R(\theta_q+\epsilon_{1n})-F_R(\theta_q-\epsilon_{2n})=f(\theta_q)(\epsilon_{1n}+\epsilon_{2n})+o_p(n^{-1/2})$$

for any $\epsilon_{jn}=O_p(n^{-1/2}), j=1,2$. Then by Theorem 2.3 and following the proof of Theorem 4 in Francisco and Fuller(1991), we have

$$F_R^{-1}(q)\pm n^{-1/2}z_{\alpha/2}s_n\{f(\theta_q)\}^{-1}=F_R^{-1}(q\pm n^{-1/2}z_{\alpha/2}s_n)+o_p(n^{-1/2}).$$

Therefore, to prove (2.14), we only need to show that

$$\begin{aligned}&P\left[F_R^{-1}(q)-z_{\alpha/2}n^{-1/2}s_n\{f(\theta_q)\}^{-1}\le\theta_q\le F_R^{-1}(q)+z_{\alpha/2}n^{-1/2}s_n\{f(\theta_q)\}^{-1}\right]\\&\to 1-\alpha,\end{aligned}$$

which is implied by Theorem 2.3.

(W2). Woodruff-type CI under adjusted random imputation:

$$\left[F_A^{-1}(q-z_{\alpha/2}n^{-1/2}\hat{\sigma}_{A1,q}),F_A^{-1}(q+z_{\alpha/2}n^{-1/2}\hat{\sigma}_{A1,q})\right].$$

We note that $n^{-1}\hat{\sigma}^2_{A1,q}$ above is a consistent estimator of the variance of $F_A(\theta_q)$, but it depends on $\hat{f}_A(\hat{\theta}_q^{(A)})$. Similar to above derivations, we can also show that the above Woodruff intervals have the asymptotically correct $(1-\alpha)$-level coverage probability.

Chen and Shao (1999) also obtained normal approximation intervals for the mean and quantiles and Woodruff intervals for quantiles under random imputation. However, they appealed to a Lemma in Schenker and Welsch (1988) that requires a stronger regularity condition than the condition 2 in Lemma 7.1 of the Appendix (Chen and Rao, 2006). We verified condition 2 explicitly in each case.

3. EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS

3.1 Mean μ

Let $Z_{m,M,i}(\mu) = Y_{M,i} - \mu,\ Z_{m,R,i}(\mu) = Y_{R,i} - \mu$, and $Z_{m,A,i}(\mu) = Y_{A,i} - \mu$. Then the empirical log-likelihood ratios for μ under the three different imputations M, R and A are defined respectively as

$$\ell_{m,M,n}(\mu) = -2 \max_{\sum_{i=1}^n p_i^{(m,M)} Z_{m,M,i}(\mu)=0,\ \sum_{i=1}^n p_i^{(m,M)}=1} \sum_{i=1}^n \log(np_i^{(m,M)}),$$

$$\ell_{m,R,n}(\mu) = -2 \max_{\sum_{i=1}^n p_i^{(m,R)} Z_{m,R,i}(\mu)=0,\ \sum_{i=1}^n p_i^{(m,R)}=1} \sum_{i=1}^n \log(np_i^{(m,R)}),$$

and

$$\ell_{m,A,n}(\mu) = -2 \max_{\sum_{i=1}^n p_i^{(m,A)} Z_{m,A,i}(\mu)=0,\ \sum_{i=1}^n p_i^{(m,A)}=1} \sum_{i=1}^n \log(np_i^{(m,A)}).$$

Note that the empirical likelihood ratios are based on the completed data $Y_{M,i},\ Y_{R,i}$ or $Y_{A,i}$, $i = 1, 2, \ldots, n$. It can be shown, by using the Lagrange multiplier method, that

$$\ell_{m,M,n}(\mu) = 2 \sum_{i=1}^n \log\{1 + \lambda_n^{(m,M)} Z_{m,M,i}(\mu)\},$$

where $\lambda_n^{(m,M)}$ is the solution of the equation

$$\frac{1}{n} \sum_{i=1}^n \frac{Z_{m,M,i}(\mu)}{1 + \lambda_n^{(m,M)} Z_{m,M,i}(\mu)} = 0,$$

$$\ell_{m,R,n}(\mu) = 2 \sum_{i=1}^n \log\{1 + \lambda_n^{(m,R)} Z_{m,R,i}(\mu)\},$$

where $\lambda_n^{(m,R)}$ is the solution of the equation

$$\frac{1}{n} \sum_{i=1}^n \frac{Z_{m,R,i}(\mu)}{1 + \lambda_n^{(m,R)} Z_{m,R,i}(\mu)} = 0,$$

and

$$\ell_{m,A,n}(\mu) = 2 \sum_{i=1}^n \log\{1 + \lambda_n^{(m,A)} Z_{m,A,i}(\mu)\},$$

where $\lambda_n^{(m,A)}$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_{m,A,i}(\mu)}{1+\lambda_n^{(m,A)}Z_{m,A,i}(\mu)}=0.$$

Results on the asymptotic distribution of the above empirical log-likelihood ratios for μ are summarized in Theorem 3.1. The proof of Theorem 3.1 is given in the Appendix.

THEOREM 3.1 *Under the conditions that* $0 < p \leq 1$ *and* $0 < \mathit{Var}(Y) < \infty$,

$$\ell_{m,M,n}(\mu) \xrightarrow{d} p^{-2}\chi_1^2 \tag{3.1}$$

as $n \to \infty$. *Further, assume that there exists an* $\alpha_0 > 0$ *such that* $E|Y|^{2+\alpha_0} < \infty$. *Then, as* $n \to \infty$,

$$\ell_{m,R,n}(\mu) \xrightarrow{d} (1-p+p^{-1})\chi_1^2 \tag{3.2}$$

and

$$\ell_{m,A,n}(\mu) \xrightarrow{d} p^{-1}\chi_1^2. \tag{3.3}$$

Using Theorem 3.1, asymptotically correct $(1-\alpha)$-level empirical likelihood based confidence intervals on μ are obtained as follows. Let $\chi_{1,\alpha}^2$ be the upper α-point of χ_1^2 variable, i.e. $P(\chi_1^2 > \chi_{1,\alpha}^2) = \alpha$. Then

(1). CI under mean imputation:

$$\{\tilde{\mu} : \hat{p}^2\ell_{m,M,n}(\tilde{\mu}) \leq \chi_{1,\alpha}^2\},$$

(2). CI under random imputation:

$$\{\tilde{\mu} : (1-\hat{p}+\hat{p}^{-1})^{-1}\ell_{m,R,n}(\tilde{\mu}) \leq \chi_{1,\alpha}^2\},$$

and

(3). CI under adjusted random imputation:

$$\{\tilde{\mu} : \hat{p}\; \ell_{m,A,n}(\tilde{\mu}) \le \chi^2_{1,\alpha}\}.$$

It follows from (1)–(3) that the EL intervals for μ depend only on the completed data and the response rate $\hat{p}$ reported in the data file. Standard EL methods for the complete response case can be applied to the data file to calculate the empirical log-likelihood ratios and hence EL intervals using (1)–(3) above.

3.2 Distribution Function θ

Let $Z_{d,R,i}(\theta) = I(Y_{R,i} \le y) - \theta$ and $Z_{d,A,i}(\theta) = I(Y_{A,i} \le y) - \theta$. Then the empirical log-likelihood ratios for θ under imputations R and A are defined respectively as

$$\ell_{d,R,n}(\theta) = -2 \max_{\sum_{i=1}^n p_i^{(d,R)} Z_{d,R,i}(\theta)=0, \sum_{i=1}^n p_i^{(d,R)}=1} \sum_{i=1}^n \log(np_i^{(d,R)}),$$

and

$$\ell_{d,A,n}(\theta) = -2 \max_{\sum_{i=1}^n p_i^{(d,A)} Z_{d,A,i}(\theta)=0, \sum_{i=1}^n p_i^{(d,A)}=1} \sum_{i=1}^n \log(np_i^{(d,A)}).$$

Again, the empirical likelihood ratios depend only on the completed data. It can be shown, by using the Lagrange multiplier method, that

$$\ell_{d,R,n}(\theta) = 2\sum_{i=1}^n \log\{1 + \lambda_n^{(d,R)} Z_{d,R,i}(\theta)\},$$

where $\lambda_n^{(d,R)}$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_{d,R,i}(\theta)}{1 + \lambda_n^{(d,R)} Z_{d,R,i}(\theta)} = 0,$$

and

$$\ell_{d,A,n}(\theta) = 2\sum_{i=1}^n \log\{1 + \lambda_n^{(d,A)} Z_{d,A,i}(\theta)\},$$

where $\lambda_n^{(d,A)}$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_{d,A,i}(\theta)}{1 + \lambda_n^{(d,A)} Z_{d,A,i}(\theta)} = 0.$$

Results on the asymptotic distribution of the above empirical log-likelihood ratios for θ are summarized in Theorem 3.2. The proof of Theorem 3.2 is given in the Appendix.

THEOREM 3.2 *Assume that $F(y) > 0$, and that there exists an $\alpha_0 > 0$ such that $E|Y|^{2+\alpha_0} < \infty$. Then as $n \to \infty$,*

$$\ell_{d,R,n}(\theta) \xrightarrow{d} (1 - p + p^{-1})\chi_1^2 \tag{3.4}$$

and

$$\ell_{d,A,n}(\theta) \xrightarrow{d} [\sigma_{A,F}^2(y)/\{F(y)(1 - F(y))\}]\chi_1^2, \tag{3.5}$$

where $\sigma_{A,F}^2(y)$ is defined in Theorem 2.2.

Using Theorem 3.2, asymptotically correct $(1 - \alpha)$-level empirical likelihood based confidence intervals on θ are obtained as follows:

(1). CI under random imputation:

$$\{\tilde{\theta} : (1 - \hat{p} + \hat{p}^{-1})^{-1}\ell_{d,R,n}(\tilde{\theta}) \leq \chi_{1,\alpha}^2\},$$

and

(2). CI under adjusted random imputation:

$$\{\tilde{\theta} : [\hat{F}(y)\{1 - \hat{F}(y)\}/\hat{\sigma}_{A,F}^2(y)]\ell_{d,A,n}(\tilde{\theta}) \leq \chi_{1,\alpha}^2\},$$

where $\hat{F}(y) = F_A(y)$ and $\hat{\sigma}_A^2$ is the same as in (2.8) so that they are consistent estimators of the corresponding population quantities.

It follows from (1) and (2) that the EL intervals for $F(y)$ depend only on the completed data and the response rate $\hat{p}$.

3.3 q-th Quantile θ_q

Let $Z_{q,R,i}(\theta_q) = I(Y_{R,i} \le \theta_q) - q$ and $Z_{q,A,i}(\theta_q) = I(Y_{A,i} \le \theta_q) - q$. Then the empirical log-likelihood ratios for $\tilde{\theta}_q$ under imputations R and A are defined respectively as

$$\ell_{q,R,n}(\theta_q) = -2 \max_{\sum_{i=1}^n p_i^{(q,R)} Z_{q,R,i}(\theta_q)=0, \sum_{i=1}^n p_i^{(q,R)}=1} \sum_{i=1}^n \log(np_i^{(q,R)}),$$

and

$$\ell_{q,A,n}(\theta_q) = -2 \max_{\sum_{i=1}^n p_i^{(q,A)} Z_{q,A,i}(\theta_q)=0, \sum_{i=1}^n p_i^{(q,A)}=1} \sum_{i=1}^n \log(np_i^{(q,A)}).$$

Again, the empirical likelihood ratios depend only on the completed data. It can be shown, by using the Lagrange multiplier method, that

$$\ell_{q,R,n}(\theta_q) = 2\sum_{i=1}^n \log\{1 + \lambda_n^{(q,R)} Z_{q,R,i}(\theta_q)\},$$

where $\lambda_n^{(q,R)}$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_{q,R,i}(\theta_q)}{1 + \lambda_n^{(q,R)} Z_{q,R,i}(\theta_q)} = 0,$$

and

$$\ell_{q,A,n}(\theta_q) = 2\sum_{i=1}^n \log\{1 + \lambda_n^{(q,A)} Z_{q,A,i}(\theta_q)\},$$

where $\lambda_n^{(q,A)}$ is the solution of the equation

$$\frac{1}{n}\sum_{i=1}^n \frac{Z_{q,A,i}(\theta_q)}{1 + \lambda_n^{(q,A)} Z_{q,A,i}(\theta_q)} = 0.$$

Results on the asymptotic distribution of the above empirical log-likelihood ratios for θ_q are summarized in Theorem 3.3. The proof of Theorem 3.3 is given in the Appendix.

THEOREM 3.3 *Under conditions of Theorem 2.3, as $n \to \infty$,*

$$\ell_{q,R,n}(\theta_q) \xrightarrow{d} (1 - p + p^{-1})\chi_1^2, \tag{3.6}$$

and

$$\ell_{q,A,n}(\theta_q) \xrightarrow{d} [\sigma_{A1,q}^2 \{q(1-q)\}^{-1}]\chi_1^2, \tag{3.7}$$

where $\sigma_{A1,q}^2$ is defined in Theorem 2.3.

Using Theorem 3.3, asymptotically correct $(1-\alpha)$–level empirical likelihood based confidence intervals on θ_q are obtained as follows:

(1). CI under random imputation:

$$\{\tilde{\theta}_q : (1 - \hat{p} + \hat{p}^{-1})^{-1}\ell_{q,R,n}(\tilde{\theta}_q) \leq \chi_{1,\alpha}^2\},$$

and

(2). CI under adjusted random imputation:

$$\{\tilde{\theta}_q : \{q(1-q)/\hat{\sigma}_{A1,q}^2\}\ell_{q,A,n}(\tilde{\theta}_q) \leq \chi_{1,\alpha}^2\},$$

where $\hat{\sigma}_{A1,q}^2$ is the same as in (2.13), which is a consistent estimator of $\sigma_{A1,q}^2$. It follows from (1) and (2) that the EL intervals for θ_q depend only on the completed data and the response rate $\hat{p}$.

4. SIMULATION STUDY

We conducted a small simulation study on the finite sample performance of normal approximation and empirical likelihood based confidence intervals on the mean $\mu = E(Y)$, distribution function $\theta = F(y)$ for fixed y and quantile $\theta_q = F^{-1}(q)$. Random samples $\{Y_i,\ \delta_i;\ i = 1, \ldots, n\}$ were generated from the standard exponential distribution with mean 1 and three cases of uniform response probabilities, $p = 0.7,\ 0.8,\ 0.9$.

For each of the three cases, we generated $10,000$ random samples of incomplete data $\{Y_i, \delta_i,\ i = 1, \cdots, n\}$ for $n = 60$ and 120. For nominal confidence level of 95%, using the simulated samples, we evaluated the coverage probability (CP), lower tail error rate (L), upper tail error rate (U) and the average length of the interval (AL) of the normal approximation based (NA) and empirical likelihood based (EL) intervals for the three imputation methods: mean imputation (M), random hot deck imputation (R) and adjusted random hot deck imputation (A). In the case of quantiles, we denote the Woodruff type confidence intervals as W.

Table 1 reports the simulation results for the mean $\mu = E(Y)$. It is seen from Table 1 that EL provides more balanced error rates (L and U) than NA under the three different imputation methods M, R and A. In the case of NA, L is significantly lower and U is significantly higher than the nominal 2.5%. For example, for $n = 60$, $p = 0.7$ and mean imputation (M), L$=$ 1.3% and U$=$ 6.0% for NA compared to L$=$ 3.0% and U$=$ 3.4% for EL. The imbalance in error rates decreases as n increases. Under M and R, the performance of EL in terms of CP is slightly better than NA, but NA seems to be slightly better than EL under A. In terms of average length (AL), M and A perform similarly whereas R leads to larger AL, as expected. Also, NA performs slightly better than EL in terms of AL but at the expense of undercoverage.

Table 2 reports the simulation results for the distribution function $\theta = F(y) = 0.86$ under R and A; note that M is not suitable for θ and the quantile θ_q since it leads to asymptotically inconsistent imputed estimators. It is clear from Table 2 that EL outperforms NA in terms of CP, with values closer to nominal 95% even for $n = 60$, and balanced error rates L and U. For example, with $n = 60$, $p = 0.7$ and random imputation R, CP$=$ 90.9%, L$=$ 7.9% and U $=$ 1.2% for NA compared to CP $=$ 95.1%, L $=$ 2.3% and U$=$ 2.6% for EL. Again, NA is better than EL in terms of AL but at the expense of undercoverage.

Table 3 reports the simulation results for the median $\theta_{\frac{1}{2}} = F^{-1}(\frac{1}{2})$. Here NA leads to severe undercoverage whereas the Woodruff (W) method of EL leads to CP closer to nominal 95%. For example, with $n = 60$, $p = 0.7$, and random imputation (R), CP$=$ 87.3% for NA compared to CP$=$ 95.0% for EL, and CP $=$ 95.5% for W. Also, EL and W provide similar results in terms of CP, L, U and AL, although AL is slightly smaller for EL. Our results suggest that NA is not recommended for quantiles, and either EL or W should be used in practice. However, EL provides a unified method for all the parameters μ, θ and θ_q whereas W is tailor-made for θ_q.

5. STRATIFIED RANDOM SAMPLING

5.1 Normal approximation intervals

Suppose that the population is divided into H strata with known relative sizes W_h, $h = 1, \ldots, H$; $\sum_{h=1}^{H} W_h = 1$. Independent simple random samples of sizes n_h, $h = 1, \ldots, H$ are drawn from the strata, and the strata sampling fractions, n_h/N_h, are assumed to be negligible. We express μ, θ and θ_q as $\mu = \Sigma W_h \mu_h$, $F(y) = \Sigma W_h F_h(y)$ and $\theta_q = F^{-1}(q)$. We regard the sample of incomplete data in stratum h, $\{(y_{hi}, \delta_{hi}),\ i = 1, \ldots, n_h\}$ as an i.i.d. sample generated from the random vector (Y_h, δ_h). Put $n = \sum_h n_h$. We assume MCAR mechanism within each stratum, i.e., $P(\delta_h = 1|Y_h) = P(\delta_h = 1) = p_h$, $0 < p_h \leq 1$. Imputations M, R or A are performed separately in each stratum, and we have $\bar{Y}_M = \Sigma W_h \bar{Y}_{Mh}$, $\bar{Y}_R = \Sigma W_h \bar{Y}_{Rh}$ and $\bar{Y}_A = \Sigma W_h \bar{Y}_{Ah}$ as estimators of μ. We obtain an extension of Theorem 2.1 by letting $n_h \to \infty$ for each h with fixed H and assuming that $n/n_h \to \lambda_h (0 < \lambda_h < \infty)$ and that $0 < \mathrm{var}(Y_h) = \sigma_h^2 < \infty$. We assume that the imputed data file provides stratum identifiers and stratum response rates $\hat{p}_h = r_h/n_h$. Identification flags on the imputed values are not needed.

Normal approximation based $(1-\alpha)$–level intervals on μ are given by $\bar{Y}_M \pm z_{\alpha/2}[\Sigma W_h^2 n_h^{-1} \hat{p}_h^{-1} s_{Mh}^2]^{1/2}$, $\bar{Y}_R \pm z_{\alpha/2}[\Sigma W_h^2 n_h^{-1}(1-\hat{p}_h+\hat{p}_h^{-1}) s_{Rh}^2]^{1/2}$ and $\bar{Y}_A \pm z_{\alpha/2}[\Sigma W_h^2 n_h^{-1} \hat{p}_h^{-1} s_{Ah}^2]^{1/2}$ under M, R and A respectively, using obvious extension of the notation for simple random sampling.

Estimators of $F(y)$ under R and A are given by $F_R(y) = \Sigma W_h F_{Rh}(y)$ and $F_A(y) = \Sigma W_h F_{Ah}(y)$. Normal approximation based $(1-\alpha)$-level intervals are given by $F_R(y) \pm z_{\alpha/2}[\Sigma W_h^2 n_h^{-1}(1-\hat{p}_h+\hat{p}_n^{-1})\hat{\sigma}^2_{R,F_h}(y)]^{1/2}$ and $F_A(y) \pm z_{\alpha/2}[\Sigma W_h^2 n_h^{-1} \hat{\sigma}^2_{A,F_h}(y)]^{1/2}$ under R and A respectively, using obvious extension of previous notation for simple random sampling. Specifically, $\hat{\sigma}^2_{R,F_h}$ is the estimator of $F_h(y)\{1-F_h(y)\}$.

We focus only on the Woodruff intervals for quantiles under R and A because normal approximation based intervals for quantiles did not perform well under simple random sampling in our simulation study (Section 4). The $(1-\alpha)$–level Woodruff intervals on θ_q under R and A are given by

$$\Big[F_R^{-1}\left(q - z_{\alpha/2}\{\Sigma W_h^2 n_h^{-1}(1-\hat{p}_h+\hat{p}_h^{-1})\hat{\sigma}^2_{R,F_h}(\hat{\theta}_q)\}^{1/2}\right), \\ F_R^{-1}\left(q + z_{\alpha/2}\{\Sigma W_h^2 n_h^{-1}(1-\hat{p}_h+\hat{p}_h^{-1})\hat{\sigma}^2_{R,F_h}(\hat{\theta}_q)\}^{1/2}\right)\Big]$$

and

$$\Big[F_A^{-1}\Big(q - z_{\alpha/2}\{\Sigma W_h^2 n_h^{-1}\hat{\sigma}^2_{A1,h,q}\}^{1/2}\Big),\quad F_A^{-1}\Big(q + z_{\alpha/2}\{\Sigma W_h^2 n_h^{-1}\hat{\sigma}^2_{A1,h,q}\}^{1/2}\Big)\Big]$$

respectively, using obvious extension of previous notation for simple random sampling.

5.2 EL intervals

We now obtain EL intervals under stratified random sampling. For EL based CI on μ, under M, we maximize $\Sigma_h\Sigma_i \log(n_h p_{hi}^{(m,M)})$ subject to $\Sigma_i p_{hi}^{(m,M)} = 1$,

$h = 1,\ldots,H$ and $\Sigma_h W_h \Sigma_i p_{hi}^{(m,M)} Y_{M,hi} = \mu$, leading to empirical log-likelihood ratio

$$\begin{aligned}\ell_{m,M,\mathbf{n}}(\mu) &= -2 \max_{p_{hi},1\le h\le H,1\le i\le n_h} \Sigma_h\Sigma_i \log(n_h p_{hi}^{(m,M)}) \\ &= 2\Sigma_h\Sigma_i \log\{1+m_h t(\mu)(Y_{M,hi}-\psi_{m,M,h}(\mu))\},\end{aligned}$$

where $\mathbf{n} = (n_1,\ldots,n_H)', m_h = nW_h n_h^{-1}$, and $\psi_{m,M,h}(\mu), t(\mu)$ satisfy

$$\begin{cases} \sum_i \frac{Y_{M,hi}-\psi_{m,M,h}(\mu)}{1+m_h t(\mu)(Y_{M,hi}-\psi_{m,M,h}(\mu))} = 0, 1\le h\le H, \\ \sum_h W_h\psi_{m,M,h}(\mu) = \mu. \end{cases} \tag{5.1}$$

Zhong and Rao (2000) and Wu (2004) have given algorithms for evaluating empirical log-likelihood ratio for the complete data case. Here the same algorithms can be applied to the imputed data file to calculate $\ell_{m,M,\mathbf{n}}(\mu)$. Similarly, $\ell_{m,R,\mathbf{n}}(\mu)$ and $\ell_{m,A,\mathbf{n}}(\mu)$ are obtained. It can be shown, under the assumption that $\hat\lambda_h = n/n_h \to \lambda_h(0<\lambda_h<\infty)$, that $\ell_{m,M,\mathbf{n}}(\mu)$, $\ell_{m,R,\mathbf{n}}(\mu)$ and $\ell_{m,A,\mathbf{n}}(\mu)$ respectively have limiting distributions

$$\sum_h W_h^2\lambda_h p_h^{-1}\sigma_h^2(\sum_h W_h^2\lambda_h\sigma_{M,h}^2)^{-1}\chi_1^2,\ \sum_h W_h^2\lambda_h(1-p_h+p_h^{-1})\sigma_h^2(\sum_h W_h^2\lambda_h\sigma_{R,h}^2)^{-1}\chi_1^2$$

and $\sum_h W_h^2\lambda_h p_h^{-1}\sigma_h^2(\sum_h W_h^2\lambda_h\sigma_{A,h}^2)^{-1}\chi_1^2$ respectively, where $\sigma_{M,h}^2 = p_h\sigma_h^2+(\mu_h-\psi_{m,M,h}(\mu))^2$, $\sigma_{R,h}^2 = \sigma_h^2+(\mu_h-\psi_{m,R,h}(\mu))^2$ and $\sigma_{A,h}^2 = \sigma_h^2+(\mu_h-\psi_{m,A,h}(\mu))^2$. Thus EL based $(1-\alpha)$–level intervals on μ are given by

$$\Big\{\tilde\mu : \Big(\sum_h W_h^2\hat\lambda_h\hat p_h^{-1}s_{Mh}^2\Big)^{-1}\Big[\sum_h W_h^2\hat\lambda_h\hat\sigma_{M,h}^2(\tilde\mu)\Big]\ell_{m,M,\mathbf{n}}(\tilde\mu)\le\chi_{1,\alpha}^2\Big\},$$

$$\Big\{\tilde\mu : \Big(\sum_h W_h^2\hat\lambda_h(1-\hat p_h+\hat p_h^{-1})s_{Rh}^2\Big)^{-1}\Big[\sum_h W_h^2\hat\lambda_h\hat\sigma_{R,h}^2(\tilde\mu)\Big]\ell_{m,R,\mathbf{n}}(\tilde\mu)\le\chi_{1,\alpha}^2\Big\}$$

and

$$\Big\{\tilde\mu : \Big(\sum_h W_h^2\hat\lambda_h\hat p_h^{-1}s_{Ah}^2\Big)^{-1}\Big[\sum_h W_h^2\hat\lambda_h\hat\sigma_{A,h}^2(\tilde\mu)\Big]\ell_{m,A,\mathbf{n}}(\tilde\mu)\le\chi_{1,\alpha}^2\Big\}$$

under M, R and A respectively, using obvious extension of the notation for simple random sampling, where $\hat\sigma_{M,h}^2(\mu) = \hat p_h s_{Mh}^2+(\bar Y_{Mh}-\psi_{m,M,h}(\mu))^2$, $\hat\sigma_{R,h}^2(\mu) = s_{Rh}^2+(\bar Y_{Rh}-\psi_{M,R,h}(\mu))^2$ and $\hat\sigma_{A,h}^2(\mu) = s_{Ah}^2+(\bar Y_{Ah}-\psi_{m,A,h}(\mu)^2$.

We now turn to EL intervals on $\theta = F(y)$ under R and A. Under R, the empirical log-likelihood ratio $\ell_{d,R,\mathbf{n}}(\theta) = -2\max_{p_{hi},1\leq h\leq H,1\leq i\leq n_h} \Sigma_h\Sigma_i \log(n_h p_{hi}^{(d,R)})$ subject to $\Sigma_i p_{hi}^{(d,R)} = 1$, $h = 1,\ldots,H$ and $\Sigma_h W_h \Sigma_i p_{hi}^{(d,R)} I(Y_{R,hi} \leq y) = \theta$. $\ell_{d,A,\mathbf{n}}(\theta)$ is obtained using $I(Y_{A,hi} \leq y)$ similarly. The EL based $(1-\alpha)$–level intervals on θ are given by

$$\Big\{\tilde{\theta} : \Big(\sum_h W_h^2 \hat{\lambda}_h (1-\hat{p}_h+\hat{p}_h^{-1})\hat{\sigma}^2_{R,F_h}(y)\Big)^{-1} \times\Big[\sum_h W_h^2\hat{\lambda}_h(\hat{\sigma}^2_{R,F_h}(y)+\hat{\Delta}^2_{R,h}(\tilde{\theta}))\Big]\ell_{d,R,\mathbf{n}}(\tilde{\theta}) \leq \chi^2_{1,\alpha}\Big\}$$

and

$$\Big\{\tilde{\theta} : \Big(\sum_h W_h^2 \hat{\lambda}_h \hat{\sigma}^2_{A,F_h}(y)\Big)^{-1}\Big[\sum_h W_h^2\hat{\lambda}_h(\hat{\sigma}^2_{0,A,F_h}(y)+\hat{\Delta}^2_{A,h}(\tilde{\theta}))\Big]\ell_{d,A,\mathbf{n}}(\tilde{\theta}) \leq \chi^2_{1,\alpha}\Big\}$$

under R and A respectively, using obvious extension of the notation for simple random sampling, where $\hat{\sigma}^2_{0,A,F_h}(y) = F_{Ah}(y)\{1-F_{Ah}(y)\}$, $\hat{\Delta}_{R,h}(\tilde{\theta}) = F_{Rh}(y) - \psi_{d,R,h}(\tilde{\theta})$ and $\hat{\Delta}_{A,h}(\tilde{\theta}) = F_{Ah}(y) - \psi_{d,A,h}(\tilde{\theta})$ respectively.

Finally, we investigate the EL based CI on $\theta_q = F^{-1}(q)$. Under R, the empirical log-likelihood ratio $\ell_{q,R,\mathbf{n}}(\theta_q) = -2\max_{p_{hi},1\leq h\leq H,1\leq i\leq n_h} \Sigma_h\Sigma_i \log(n_h p_{hi}^{(q,R)})$ subject to $\Sigma_i p_{hi}^{(q,R)} = 1$, $h = 1,\ldots,H$ and $\Sigma_h W_h \Sigma_i p_{hi}^{(q,R)} Z_{q,R,hi} I(Y_{R,hi} \leq \theta_q) = q$. $\ell_{q,A,\mathbf{n}}(\theta_q)$ is obtained similarly. The EL based $(1-\alpha)$–level intervals on θ_q are given by

$$\Big\{\tilde{\theta}_q : \Big(\sum_h W_h^2 \hat{\lambda}_h (1-\hat{p}_h+\hat{p}_h^{-1})q(1-q)\Big)^{-1} \times\Big[\sum_h W_h^2\hat{\lambda}_h(q(1-q)+\hat{\Delta}^2_{q,R,h}(\tilde{\theta}_q))\Big]\ell_{q,R,\mathbf{n}}(\tilde{\theta}_q) \leq \chi^2_{1,\alpha}\Big\}$$

and

$$\Big\{\tilde{\theta}_q : \Big(\sum_h W_h^2 \hat{\lambda}_h \hat{\sigma}^2_{A1,h,q}\Big)^{-1}\Big[\sum_h W_h^2\hat{\lambda}_h(q(1-q)+\hat{\Delta}^2_{q,A,h}(\tilde{\theta}_q))\Big]\ell_{q,A,\mathbf{n}}(\tilde{\theta}_q) \leq \chi^2_{1,\alpha}\Big\}$$

under R and A respectively, using obvious extension of the notation for simple random sampling, where $\hat{\Delta}_{q,R,h}(\tilde{\theta}_q) = F_{Rh}(\tilde{\theta}_q) - \psi_{q,R,h}(\tilde{\theta}_q)$ and $\hat{\Delta}_{q,A,h}(\tilde{\theta}_q) = F_{Ah}(\tilde{\theta}_q) - \psi_{q,A,h}(\tilde{\theta}_q)$ respectively.

6. SUMMARY AND CONCLUSIONS

In this paper we considered three different methods of imputation to fill in the missing values in a random sample $\{Y_i, i = 1, \ldots, n\}$: mean imputation (M), random hot deck imputation (R) and adjusted random hot deck imputation (A). Assuming uniform response probability p, we have obtained asymptotically correct normal approximation (NA) based confidence intervals on the mean μ, distribution function $\theta = F(y)$ and q-th quantile $\theta_q = F^{-1}(q)$. Asymptotically correct empirical likelihood (EL) intervals are also obtained by first showing that the empirical log-likelihood ratios are asymptotically scaled χ_1^2 variables. Both NA and EL intervals do not require identification flags on the imputed values in the data file; only the estimated response rate $\hat{p}$ is needed with the imputed data file. Simulation results indicated that EL performs better than NA in providing balanced lower (L) and upper (U) tail error rates. Also, NA lead to severe undercoverage in the case of median ($\theta_{\frac{1}{2}}$) unlike EL and the method of Woodruff (1952).

If the objective is to estimate different parameters μ, θ and θ_q from the imputed data file, then mean imputation (M) is not suitable and normal approximation (NA) leads to severe undercoverage in the case of θ_q and unbalanced tail error rates in the case of μ and θ, unlike EL. We recommend the use of random (R) or adjusted random (A) imputation and EL intervals for all the parameters.

Extensions to complex sampling designs, based on the pseudo-EL approach of Chen and Sitter (1999), and multiple imputation classes are under investigation.

Acknowledgement

This research was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada awarded to J. N. K. Rao.

7. APPENDIX: PROOFS

The following lemma of Chen and Rao (2006) will be used in the proofs of main results.

LEMMA 7.1 *Let U_n, V_n be two sequences of random variables and $\mathcal{B}_n$ be a σ-algebra. Assume that:* 1. *There exists $\sigma_{1n} > 0$ such that*

$$\sigma_{1n}^{-1} V_n \xrightarrow{d} N(0,1)$$

as $n \to \infty$, and V_n is $\mathcal{B}_n$ measurable. 2. *$E[U_n|\mathcal{B}_n] = 0$ and $Var(U_n|\mathcal{B}_n) = \sigma_{2n}^2$ such that*

$$\sup_t |P(\sigma_{2n}^{-1} U_n \leq t|\mathcal{B}_n) - \Phi(t)| = o_p(1),$$

where $\Phi(\cdot)$ is the distribution function of the standard normal random variable. 3. $\gamma_n^2 = \sigma_{1n}^2/\sigma_{2n}^2 = \gamma^2 + o_p(1)$. *Then, as $n \to \infty$,*

$$\frac{U_n + V_n}{\sqrt{\sigma_{1n}^2 + \sigma_{2n}^2}} \xrightarrow{d} N(0,1).$$

PROOF OF THEOREM 2.1. Noting that $\bar{Y}_M = \bar{Y}_A = \bar{Y}_r$, it follows that

$$\begin{aligned}
\sqrt{n}(\bar{Y}_M - \mu) = \sqrt{n}(\bar{Y}_A - \mu) &= \sqrt{n}\Big\{\frac{1}{r}\sum_{i \in s_r}(Y_i - \mu)\Big\} \\
&= \sqrt{n}\Big\{\frac{1}{r}\sum_{i=1}^{n}\delta_i(Y_i - \mu)\Big\} = \frac{n}{\sum_{i=1}^n \delta_i}\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i(Y_i - \mu)\Big\} \\
&= \{p + o_p(1)\}^{-1}\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i(Y_i - \mu)\Big\}.
\end{aligned}$$

So from the Central Limit Theorem for i.i.d. random variables and MCAR assumption, we have (2.1) and (2.2), noting that $E(\delta_i|Y_i) = E(\delta_i) = p$ and $V(\delta_i|Y_i) = E(\delta_i|Y_i) - [E(\delta_i|Y_i)]^2 = p - p^2$. Let $V_n = \sqrt{n}(\bar{Y}_M - \mu), U_n = \sqrt{n}(\bar{Y}_R - \bar{Y}_M)$ and $\mathcal{B}_n = \sigma((\delta_i, Y_i), i = 1, \cdots, n)$. So V_n is $\mathcal{B}_n$ measurable, and $\sqrt{n}(\bar{Y}_R - \mu) = V_n + U_n$. If we let $\sigma_{1n}^2 = p^{-1}\sigma^2$, then from (2.1),

$$\sigma_{1n}^{-1} V_n \xrightarrow{d} N(0,1). \tag{7.1}$$

We now verify condition 2 in Lemma 7.1. It can be seen, for $i \in s_m$, that

$$E(Y_i^{(R)}|\mathcal{B}_n) = \bar{Y}_r, \mathrm{Var}(Y_i^{(R)}|\mathcal{B}_n) = \frac{1}{r}\sum_{i\in s_r}(Y_i - \bar{Y}_r)^2.$$

It follows that

$$E(\bar{Y}_R|\mathcal{B}_n) = \bar{Y}_r, \mathrm{Var}(\bar{Y}_R|\mathcal{B}_n) = \frac{n-r}{n^2}\left\{\frac{1}{r}\sum_{i\in s_r}(Y_i - \bar{Y}_r)^2\right\}.$$

Let $\sigma_{2n}^2 = \frac{n-r}{n}\left\{\frac{1}{r}\sum_{i\in s_r}(Y_i - \bar{Y}_r)^2\right\}$. Then

$$E(U_n|\mathcal{B}_n) = 0, \mathrm{Var}(U_n|\mathcal{B}_n) = \sigma_{2n}^2.$$

Similar to the proof of (2.1), it can be shown that $\sigma_{2n}^2 = (1-p)\sigma^2 + o_p(1)$. Further,

$$\begin{aligned}
&\frac{\sum_{i=1}^n E(|Y_{R,i}|^{2+\alpha_0}|\mathcal{B}_n)}{\{\sum_{i=1}^n E(Y_{R,i}^2|\mathcal{B}_n)\}^{(2+\alpha_0)/2}} \\
&= \frac{n^{-1-\frac{\alpha_0}{2}}r^{-1}\sum_{i=1}^n\sum_{j\in s_r}|\delta_iY_i + (1-\delta_i)Y_j|^{2+\alpha_0}}{\{n^{-1}r^{-1}\sum_{i=1}^n\sum_{j\in s_r}(\delta_iY_i + (1-\delta_i)Y_j)^2\}^{(2+\alpha_0)/2}}.
\end{aligned} \tag{7.2}$$

It is clear that

$$n^{-1}r^{-1}\sum_{i=1}^n\sum_{j\in s_r}(\delta_iY_i + (1-\delta_i)Y_j)^2 = \frac{1}{r}\sum_{i=1}^n\delta_iY_i^2 = \sigma^2 + \mu^2 + o_p(1).$$

On the other hand, there is a constant C_0 depending only on α_0 such that

$$\begin{aligned}
&n^{-1}r^{-1}\sum_{i=1}^n\sum_{j\in s_r}|\delta_iY_i + (1-\delta_i)Y_j|^{2+\alpha_0} \\
&\quad\le C_0n^{-1}r^{-1}\sum_{i=1}^n\sum_{j\in s_r}(|Y_i|^{2+\alpha_0} + |Y_j|^{2+\alpha_0}) = 2C_0E|Y|^{2+\alpha_0} + o_p(1).
\end{aligned}$$

It follows that the right hand of (7.2) converges to 0 in probability. So by Berry-Esseen's Central Limit Theorem, $\sup_t |P(\sigma_{2n}^{-1}U_n \le t|\mathcal{B}_n) - \Phi(t)| = o_p(1)$. Hence, (2.3) follows from Lemma 7.1, and the proof of Theorem 2.1 is complete.

Proof of Theorem 2.2. Denote $\bar{F}_r(y) = \frac{1}{r}\sum_{i\in s_r} I(Y_i \le y)$. Then $\sqrt{n}(\bar{F}_r(y) - \theta) = \frac{n}{r}\frac{1}{\sqrt{n}}\sum_{i=1}^n \delta_i\{I(Y_i \le y) - \theta\}$. So from the Central Limit Theorem and MCAR assumption, we have

$$[p^{-1}F(y)\{1-F(y)\}]^{-1/2}\sqrt{n}(\bar{F}_r(y)-\theta) \xrightarrow{d} N(0,1). \tag{7.3}$$

Let $V_n = \sqrt{n}(\bar{F}_r(y)-\theta), U_n = \sqrt{n}(\bar{F}_R(y)-\bar{F}_r(y))$ and $\mathcal{B}_n = \sigma((\delta_i, Y_i), i = 1,\cdots,n)$. So V_n is $\mathcal{B}_n$ measurable, and $\sqrt{n}(\bar{F}_R(y)-\theta) = V_n + U_n$. We note that

$$\begin{aligned}\bar{F}_R(y) &= \frac{1}{n}\sum_{i=1}^{n} I(Y_{R,i} \le y, \delta_i = 1) + \frac{1}{n}\sum_{i=1}^{n} I(Y_{R,i} \le y, \delta_i = 0)\\ &= \frac{r}{n}\bar{F}_r(y) + \frac{1}{n}\sum_{i\in s_m} I(Y_i^{(R)} \le y).\end{aligned}$$

So

$$\begin{aligned}U_n &= \frac{1}{\sqrt{n}}\sum_{i\in s_m}\{I(Y_i^{(R)} \le y) - \bar{F}_r(y)\}\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\{I(Y_i^{(R)} \le y) - \bar{F}_r(y)\}.\end{aligned}$$

We now verify condition 2 in Lemma 7.1. It can be seen, for $i \in s_m$, that

$$E(I(Y_i^{(R)} \le y)|\mathcal{B}_n) = \bar{F}_r(y), \mathrm{Var}(I(Y_i^{(R)} \le y)|\mathcal{B}_n) = \bar{F}_r(y)\{1-\bar{F}_r(y)\}.$$

It follows that

$$E(U_n|\mathcal{B}_n) = 0, \mathrm{Var}(U_n|\mathcal{B}_n) = \frac{n-r}{n}\Big[\bar{F}_r(y)\{1-\bar{F}_r(y)\}\Big].$$

Let

$$\sigma_{2n}^2 = \frac{n-r}{n}\Big[\bar{F}_r(y)\{1-\bar{F}_r(y)\}\Big].$$

Then $\mathrm{Var}(U_n|\mathcal{B}_n) = \sigma_{2n}^2$. It can be shown that

$$\sigma_{2n}^2 = (1-p)F(y)\{1-F(y)\} + o_p(1).$$

So by Berry-Esseen's Central Limit Theorem, $\sup_t |P(\sigma_{2n}^{-1}U_n \le t|\mathcal{B}_n) - \Phi(t)| = o_p(1)$. Hence, (2.6) follows from Lemma 7.1. Denote $F_m^*(y) = \frac{1}{m}\sum_{i\in s_m} I(Y_i^{(R)} \le y)$ and $a_n = \bar{Y}_m^{(R)} - \bar{Y}_r$. Similar to the proof of Lemma 1 in Chen and Shao (1999), it can be shown that

$$\{F_m^*(y+a_n) - F_m^*(y)\} - \{\bar{F}_r(y+a_n) - \bar{F}_r(y)\} = o_p(n^{-1/2}), \tag{7.4}$$

and

$$\{\bar{F}_r(y+a_n)-\bar{F}_r(y)\}-\{F(y+a_n)-F(y)\}=o_p(n^{-1/2}). \tag{7.5}$$

Thus, from (7.4), (7.5) and the conditions in Theorem 2.2,

$$\begin{aligned} F_A(y) &= \frac{r}{n}\bar{F}_r(y)+\frac{m}{n}F_m^*(y+a_n) \\ &= \frac{r}{n}\bar{F}_r(y)+\frac{m}{n}F_m^*(y)+\frac{m}{n}\{F_m^*(y+a_n)-F_m^*(y)\} \\ &= \frac{r}{n}\bar{F}_r(y)+\frac{m}{n}F_m^*(y)+\frac{m}{n}\{\bar{F}_r(y+a_n)-\bar{F}_r(y)+o_p(n^{-1/2})\} \\ &= \frac{r}{n}\bar{F}_r(y)+\frac{m}{n}F_m^*(y)+\frac{m}{n}\{F(y+a_n)-F(y)+o_p(n^{-1/2})\} \\ &= \frac{r}{n}\bar{F}_r(y)+\frac{m}{n}F_m^*(y)+\frac{m}{n}\{f(y)a_n+o_p(n^{-1/2})\} \\ &= \{\frac{r}{n}\bar{F}_r(y)-\frac{m}{n}f(y)\bar{Y}_r\}+\frac{m}{n}\{F_m^*(y)+f(y)\bar{Y}_m^{(R)}\}+o_p(n^{-1/2}) \\ &= \bar{F}_r(y)+\frac{m}{n}[F_m^*(y)+f(y)\bar{Y}_m^{(R)}-\{\bar{F}_r(y)+f(y)\bar{Y}_r\}] \\ &\quad +o_p(n^{-1/2}). \end{aligned} \tag{7.6}$$

To prove (2.7), let $V_n=\sqrt{n}(\bar{F}_r(y)-\theta), U_n=\sqrt{n}\frac{m}{n}[F_m^*(y)+f(y)\bar{Y}_m^{(R)}-\{\bar{F}_r(y)+f(y)\bar{Y}_r\}]$ and $\mathcal{B}_n=\sigma((\delta_i,Y_i),i=1,\cdots,n)$. So V_n is $\mathcal{B}_n$ measurable, and $\sqrt{n}(\bar{F}_A(y)-\theta)=V_n+U_n$. We now verify condition 2 in Lemma 7.1. It can be seen, for $i\in s_m$, that

$$E(\{I(Y_i^{(R)}\le y)+f(y)Y_i^{(R)}\}|\mathcal{B}_n)=\bar{F}_r(y)+f(y)\bar{Y}_r,$$

$$\begin{aligned} &\mathrm{Var}(\{I(Y_i^{(R)}\le y)+f(y)Y_i^{(R)}\}|\mathcal{B}_n) \\ &=\frac{1}{r}\sum_{i\in s_r}\{I(Y_i\le y)+f(y)Y_i\}^2-\{\bar{F}_r(y)+f(y)\bar{Y}_r\}^2. \end{aligned}$$

Thus, $E(U_n|\mathcal{B}_n)=0$, and

$$\begin{aligned} \mathrm{Var}(U_n|\mathcal{B}_n) &= n\cdot\frac{m^2}{n^2}\cdot\frac{1}{m}\mathrm{Var}\Big(\{I(Y_i^{(R)}\le y)+f(y)Y_i^{(R)}\}|\mathcal{B}_n\Big) \\ &= \frac{m}{n}\Big[\frac{1}{r}\sum_{i\in s_r}\{I(Y_i\le y)+f(y)Y_i\}^2-\{\bar{F}_r(y)+f(y)\bar{Y}_r\}^2\Big]. \end{aligned}$$

Let

$$\sigma_{2n}^2 = \frac{m}{n}\Big[\frac{1}{r}\sum_{i\in s_r}\{I(Y_i \le y) + f(y)Y_i\}^2 - \{\bar{F}_r(y) + f(y)\bar{Y}_r\}^2\Big].$$

Then $\mathrm{Var}(U_n|\mathcal{B}_n) = \sigma_{2n}^2$. It can be shown that

$$\sigma_{2n}^2 = (1-p)\Big\{F(y)-F^2(y)+2f(y)E(YI(Y \le y))-2f(y)F(y)\mu+f^2(y)\sigma^2\Big\}+o_p(1).$$

So by Berry-Esseen's Central Limit Theorem, $\sup_t |P(\sigma_{2n}^{-1}U_n \le t|\mathcal{B}_n) - \Phi(t)| = o_p(1)$. Hence, (2.7) follows from Lemma 7.1, and the proof of Theorem 2.2 is complete. To prove Theorem 2.3, we need the following result, which can be proved similar to the proof of Theorem 2.2.

Lemma 7.2 *Assume that $f(\theta_q) > 0$, then for fixed $u \in R$, as $n \to \infty$,*

$$\sqrt{n}(F_R(\theta_q+n^{-1/2}\sigma_{R,q}u)-F(\theta_q+n^{-1/2}\sigma_{R,q}u)) \xrightarrow{d} N(0,(1-p+p^{-1})F(\theta_q)\{1-F(\theta_q)\}).$$

Further, assume that there exists an $\alpha_0 > 0$ such that $E|Y|^{2+\alpha_0} < \infty$, and that $f(\cdot)$ exists and continuous in a neighborhood of θ_q. Then as $n \to \infty$,

$$\sqrt{n}(F_A(\theta_q + n^{-1/2}\sigma_{A,q}u) - F(\theta_q + n^{-1/2}\sigma_{A,q}u)) \xrightarrow{d} N(0,\sigma_{A1,q}^2),$$

where $\sigma_{R,q}, \sigma_{A,q}$ and $\sigma_{A1,q}$ are defined in Theorem 2.3.

Proof of Theorem 2.3. Note that $q = F(\theta_q)$. For fixed $u \in R$, we have

$$\begin{aligned}
&P\Big\{\frac{\sqrt{n}(\hat{\theta}_q^{(R)} - \theta_q)}{\sigma_{R,q}} \le u\Big\} = P(\hat{\theta}_q^{(R)} \le \theta_q + n^{-1/2}\sigma_{R,q}u)\\
&= P\{q \le F_R(\theta_q + n^{-1/2}\sigma_{R,q}u)\}\\
&= P\Big[\sqrt{n}\{F_A(\theta_q + n^{-1/2}\sigma_{R,q}u) - F(\theta_q + n^{-1/2}\sigma_{R,q}u)\}\\
&\qquad \ge \sqrt{n}\{F(\theta_q) - F(\theta_q + n^{-1/2}\sigma_{R,q}u)\}\Big]\\
&= P\Big[\sqrt{n}\{F_A(\theta_q + n^{-1/2}\sigma_{R,q}u) - F(\theta_q + n^{-1/2}\sigma_{R,q}u)\} \ge -\sigma_{R,q}f(\theta_q)u + o(1)\Big]\\
&= P\Big[\frac{\sqrt{n}\{F_A(\theta_q + n^{-1/2}\sigma_{R,q}u) - F(\theta_q + n^{-1/2}\sigma_{R,q}u)\}}{-\sigma_{R,q}f(\theta_q)} \le u + o(1)\Big].
\end{aligned}$$

So by Lemma 7.2, we have (2.9). Similarly, we can prove (2.10). Results (2.11) and (2.12) in Theorem 2.3 can be proved similar to the proof of Theorem 2 in Chen and Shao (1999). The proof of Theorem 2.3 is thus complete.

Proof of Theorem 3.1. Similar to Owen(1990), it can be shown, under the condition $EY^2 < \infty$, that

$$\begin{aligned}&\max_{1\le i\le n} |Z_{m,M,i}(\mu)| = o_p(n^{1/2}), \max_{1\le i\le n} |Z_{m,R,i}(\mu)| = o_p(n^{1/2}),\\ &\max_{1\le i\le n} |Z_{m,A,i}(\mu)| = o_p(n^{1/2}).\end{aligned} \tag{7.7}$$

On the other hand,

$$\frac{1}{n}\sum_{i=1}^{n} Z^2_{m,M,i}(\mu) = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i(Y_i-\mu)^2 + (1-\delta_i)(\bar{Y}_r-\mu)^2\} = p\sigma^2 + o_p(1), \tag{7.8}$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z^2_{m,R,i}(\mu) &= \frac{1}{n}\sum_{i=1}^{n}\{\delta_i(Y_i-\mu)^2 + (1-\delta_i)(Y_i^{(R)}-\mu)^2\}\\ &= p\sigma^2 + o_p(1) + \frac{m}{n}\cdot\frac{1}{m}\sum_{i\in s_m}(Y_i^{(R)}-\mu)^2\\ &= p\sigma^2 + o_p(1) + (1-p)\sigma^2 + o_p(1) = \sigma^2 + o_p(1),\end{aligned} \tag{7.9}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} Z^2_{m,A,i}(\mu) = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i(Y_i-\mu)^2 + (1-\delta_i)(Y_i^{(A)}-\mu)^2\}\\ &= p\sigma^2 + o_p(1) + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)(Y_i^{(R)}-\mu+\bar{Y}_r-\bar{Y}_m^{(R)})^2\\ &= p\sigma^2 + o_p(1) + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)(Y_i^{(R)}-\mu+o_p(1))^2 = \sigma^2 + o_p(1).\end{aligned} \tag{7.10}$$

By Theorem 2.1 and (7.7) to (7.10), similar to the proof of Theorem 1 in Owen(1990) it can be shown that

$$\ell_{m,M,n}(\mu) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z^2_{m,M,i}(\mu)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{m,M,i}(\mu)\right\}^2 + o_p(1) \xrightarrow{d} p^{-2}\chi^2_1,$$

$$\begin{aligned}\ell_{m,R,n}(\mu) &= \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{m,R,i}^2(\mu)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{m,R,i}(\mu)\right\}^{2} + o_p(1)\\ &\xrightarrow{d} (1-p+p^{-1})\chi_1^2,\end{aligned}$$

and

$$\ell_{m,A,n}(\mu) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{m,A,i}^2(\mu)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{m,A,i}(\mu)\right\}^{2} + o_p(1) \xrightarrow{d} p^{-1}\chi_1^2.$$

Thus we have Theorem 3.1.

Proof of Theorem 3.2. It can be shown, by using the results in Theorem 2.2, that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} Z_{d,R,i}^2(\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n}\{I(\delta_i Y_i + (1-\delta_i)Y_i^{(R)} \le y) - 2\theta I(\delta_i Y_i + (1-\delta_i)Y_i^{(R)} \le y) + \theta^2\}\\ &= F_R(y) - 2\theta F_R(y) + \theta^2 = \theta(1-\theta) + o_p(1)\\ &= F(y)\{1-F(y)\} + o_p(1), \qquad (7.11)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} Z_{d,A,i}^2(\theta)\\ &= \frac{1}{n}\sum_{i=1}^{n}\{I(\delta_i Y_i + (1-\delta_i)Y_i^{(A)} \le y) - 2\theta I(\delta_i Y_i + (1-\delta_i)Y_i^{(A)} \le y) + \theta^2\}\\ &= F_A(y) - 2\theta F_A(y) + \theta^2 = \theta(1-\theta) + o_p(1)\\ &= F(y)\{1-F(y)\} + o_p(1), \qquad (7.12)\end{aligned}$$

By (7.11), (7.12) and the boundness of $Z_{d,R,i}(\theta)$ and $Z_{d,A,i}(\theta)$, similar to the proof of Theorem 1 in Owen (1990) it can be shown that

$$\ell_{d,R,n}(\theta) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{d,R,i}^2(\theta)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{d,R,i}(\theta)\right\}^{2} + o_p(1) \xrightarrow{d} (1-p+p^{-1})\chi_1^2,$$

and

$$\ell_{d,A,n}(\theta) = \left\{\frac{1}{n}\sum_{i=1}^{n} Z_{d,A,i}^2(\theta)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{d,A,i}(\theta)\right\}^{2} + o_p(1) \xrightarrow{d} [\sigma_A^2/\{\theta(1-\theta)\}]\chi_1^2.$$

Thus we have Theorem 3.2.

Proof of Theorem 3.3. It can be shown, by using the results in Theorem 2.2, that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Z^2_{q,R,i}(\theta_q) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{I(\delta_i Y_i + (1-\delta_i)Y_i^{(R)} \le \theta_q) - 2qI(\delta_i Y_i + (1-\delta_i)Y_i^{(R)} \le \theta_q) + q^2\} \\
&= F_R(\theta_q) - 2qF_R(\theta_q) + q^2 \\
&= F(\theta_q)\{1 - F(\theta_q)\} + o_p(1) = q(1-q) + o_(1),
\end{aligned} \tag{7.13}$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Z^2_{q,A,i}(\theta_q) \\
&= \frac{1}{n}\sum_{i=1}^{n}\{I(\delta_i Y_i + (1-\delta_i)Y_i^{(A)} \le \theta_q) - 2qI(\delta_i Y_i + (1-\delta_i)Y_i^{(A)} \le \theta_q) + q^2\} \\
&= F_A(\theta_q) - 2qF_A(\theta_q) + q^2 \\
&= F(\theta_q)\{1 - F(\theta_q)\} + o_p(1) = q(1-q) + o_(1),
\end{aligned} \tag{7.14}$$

By Theorem 2.2, (7.13), (7.14) and the boundness of $Z_{q,R,i}(\theta)$ and $Z_{q,A,i}(\theta)$, similar to the proof of Theorem 1 in Owen(1990) it can be shown that

$$\begin{aligned}
\ell_{q,R,n}(\theta) &= \left\{\frac{1}{n}\sum_{i=1}^{n} Z^2_{q,R,i}(\theta_q)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{q,R,i}(\theta_q)\right\}^2 + o_p(1) \\
&\xrightarrow{d} (1 - p + p^{-1})\chi^2_1,
\end{aligned}$$

and

$$\begin{aligned}
\ell_{q,A,n}(\theta) &= \left\{\frac{1}{n}\sum_{i=1}^{n} Z^2_{q,A,i}(\theta_q)\right\}^{-1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{q,A,i}(\theta_q)\right\}^2 + o_p(1) \\
&\xrightarrow{d} [\sigma^2_{A1,q}/\{q(1-q)\}]\chi^2_1.
\end{aligned}$$

Thus we have Theorem 3.3.

Proof of Lemma 2.1. We use the notation used in the proof of Theorem 2.2. Similar to the proof of (7.6), we have

$$\begin{aligned} F_A(y+n^{-1/2}) &= \frac{r}{n}\bar{F}_r(y+n^{-1/2}) + \frac{m}{n}F_m^*(y+n^{-1/2}+a_n) \\ &= \frac{r}{n}\bar{F}_r(y+n^{-1/2}) + \frac{m}{n}F_m^*(y) \\ &\quad + \frac{m}{n}\{f(y)(n^{-1/2}+a_n) + o_p(n^{-1/2})\}, \end{aligned}$$

and

$$\begin{aligned} F_A(y-n^{-1/2}) &= \frac{r}{n}\bar{F}_r(y-n^{-1/2}) + \frac{m}{n}F_m^*(y-n^{-1/2}+a_n) \\ &= \frac{r}{n}\bar{F}_r(y-n^{-1/2}) + \frac{m}{n}F_m^*(y) \\ &\quad + \frac{m}{n}\{f(y)(-n^{-1/2}+a_n) + o_p(n^{-1/2})\}. \end{aligned}$$

Combining with (7.5) and the conditions in this lemma, it follows that

$$\begin{aligned} &F_A(y+n^{-1/2}) - F_A(y-n^{-1/2}) \\ &= \frac{r}{n}\{\bar{F}_r(y+n^{-1/2}) - \bar{F}_r(y-n^{-1/2})\} + 2n^{-1/2}\frac{m}{n}f(y) + o_p(n^{-1/2}) \\ &= \frac{r}{n}\{F(y+n^{-1/2}) - F(y-n^{-1/2})\} + 2n^{-1/2}\frac{m}{n}f(y) + o_p(n^{-1/2}) \\ &= 2\frac{r}{n}n^{-1/2}f(y) + 2n^{-1/2}\frac{m}{n}f(y) + o_p(n^{-1/2}) = 2f(y)n^{-1/2} + o_p(n^{-1/2}). \end{aligned}$$

Thus we have Lemma 2.1.

Proof of Lemma 2.2. Let $b_n = \hat{\theta}_q^{(A)} - \theta_q$. We use the notation in the proof of Theorem 2.2. Similar to the proof of Lemma 2.1, write

$$\begin{aligned} &F_A(\hat{\theta}_q^{(A)} + n^{-1/2}) = F_A(\theta_q + b_n + n^{-1/2}) \\ &= \frac{r}{n}\bar{F}_r(\theta_q + b_n + n^{-1/2}) + \frac{m}{n}F_m^*(\theta_q + n^{-1/2} + a_n + b_n) \\ &= \frac{r}{n}\bar{F}_r(\theta_q + b_n + n^{-1/2}) + \frac{m}{n}F_m^*(\theta_q) \\ &\quad + \frac{m}{n}\{f(\theta_q)(n^{-1/2} + b_n + a_n) + o_p(n^{-1/2})\}, \end{aligned}$$

and

$$F_A(\hat{\theta}_q^{(A)} - n^{-1/2}) = F_A(\theta_q + b_n - n^{-1/2})$$

$$\begin{aligned}
&= \frac{r}{n}\bar{F}_r(\theta_q + b_n - n^{-1/2}) + \frac{m}{n}F_m^*(\theta_q - n^{-1/2} + a_n + b_n)\\
&= \frac{r}{n}\bar{F}_r(\theta_q + b_n - n^{-1/2}) + \frac{m}{n}F_m^*(\theta_q)\\
&\quad + \frac{m}{n}\{f(\theta_q)(-n^{-1/2} + b_n + a_n) + o_p(n^{-1/2})\}.
\end{aligned}$$

Combining with (7.5) and the conditions in this lemma, it follows that

$$F_A(\hat{\theta}_q^{(A)} + n^{-1/2}) - F_A(\hat{\theta}_q^{(A)} - n^{-1/2}) = 2f(\theta_q)n^{-1/2} + o_p(n^{-1/2}).$$

Thus we have the first result of Lemma 2.2. Similarly, we can prove the second result. To prove the third result, write

$$\begin{aligned}
&\frac{1}{r}\sum_{i\in s_r} Y_i I(Y_i \le \hat{\theta}_q^{(A)})\\
&= \frac{1}{r}\sum_{i\in s_r} Y_i I(Y_i \le \theta_q) + \frac{1}{r}\sum_{i\in s_r} Y_i I(\theta_q < Y_i \le \hat{\theta}_q^{(A)}, \hat{\theta}_q^{(A)} - \theta_q \ge 0)\\
&\quad + \frac{1}{r}\sum_{i\in s_r} Y_i I(\hat{\theta}_q^{(A)} < Y_i \le \theta_q, \hat{\theta}_q^{(A)} - \theta_q < 0) = I_{1n} + I_{2n} + I_{3n}. \qquad (7.15)
\end{aligned}$$

It can be shown that $I_{1n} = E\{YI(Y \le \theta_q)\} + o_p(1)$. From Theorem 2.3, $\hat{\theta}_q^{(A)} = \theta_q + o_p(1)$. So with probability tending to one, $|\hat{\theta}_q^{(A)} - \theta_q| < \delta$ for any $\delta > 0$. Thus, with probability tending to one, $I_{2n} \le \frac{1}{r}\sum_{i\in s_r}|Y_i|I(\theta_q < Y_i \le \theta_q + \delta) \le (|\theta_q| + \delta)\frac{1}{r}\sum_{i\in s_r}(\theta_q < Y_i \le \theta_q + \delta) = (|\theta_q| + \delta)\{F(\theta_q + \delta) - F(\theta_q)\} + o_p(1) = o_p(1)$ as $\delta \to 0$. Similarly, $I_{3n} = o_p(1)$. It follows that

$$\frac{1}{r}\sum_{i\in s_r} Y_i I(Y_i \le \hat{\theta}_q^{(A)}) = E\{YI(Y \le \theta_q)\} + o_p(1).$$

On the other hand, it can be shown that

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in s_m} Y_i^{(A)} I(Y_i^{(A)} \le \hat{\theta}_q^{(A)})\\
&= \frac{1}{n}\sum_{i\in s_m} Y_i^{(R)} I(Y_i^{(R)} \le \theta_q) + o_p(1) = \frac{n-r}{nr}\sum_{i\in s_r} Y_i I(Y_i \le \theta_q) + o_p(1).
\end{aligned}$$

Thus we have the third result of Lemma 2.2.

REFERENCES

Chen, J. and Rao, J. N. K., (2006). Asymptotic normality under two-phase sampling designs. *Statist. Sinica* (in press).

Chen, J., Rao, J. N. K. and Sitter, R. R., (2000). Efficient random imputation for missing data in complex surveys. *Statistica Sinica* **10**, 1153-1169.

Chen, Y. and Shao, J., (1999). Inference with survey data imputed by hot deck when imputed values are nonidentifiable. *Statist. Sinica* **9**, 361-384.

Francisco, C. A. and Fuller, W. A., (1991). Quantile estimation with a complex survey design. *Ann. Statist.* **19**, 454-469.

Hartley, H. O. and Rao, J. N. K., (1968). A new estimation theory for sample surveys. *Biometrika* **55**, 547-557.

Owen, A. B., (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-249.

Owen, A. B., (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Wang, Q. and Rao, J. N. K., (2002b). Empirical likelihood-based inference in linear models with missing data. *Scand. J. Statist.* **29**, 563-576.

Woodruff, R. S., (1952). Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.* **47**, 635-646.

Wu, C.(2004). Some Algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica.* **14** 1057-1067.

Zhong, B. and Rao, J. N. K.(2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika.* **87** 929-938.

Department of Mathematics
Guanxi Normal University
Guilin, Guanxi 541004, China
E-mail: ysqin@mailbox.gxnu.edu.cn

School of Mathematics and Statistics
Carleton University
Ottawa, ON, K1S 5B6, Canada
E-mail: qren@math.carleton.ca

Corresponding Address:

School of Mathematics and Statistics
Carleton University
Ottawa, ON, K1S 5B6, Canada
E-mail: jrao@math.carleton.ca
Tel: (+1 613)520-2600 Ext. 2146
Fax:(+1 613)520-2167

TABLE 1

Confidence interval coverage probability (CP), lower (L) and upper (U) tail error rates and average length (AL) for the mean $\mu = E(Y)$ with $p = 0.7, 0.8, 0.9$ and $n = 50, 120$: Imputation methods M, R and A; $R = 10,000$ simulations; $Y \sim \exp(1)$; NA = normal approximation, EL = empirical likelihood

			CP(%)		L(%)		U(%)		AL	
n	p	IMP	NA	EL	NA	EL	NA	EL	NA	EL
60	0.7	M	92.6	93.7	1.3	3.0	6.0	3.4	0.59	0.64
		A	92.4	91.7	1.4	3.6	6.1	4.7	0.58	0.60
		R	92.7	93.7	1.2	2.4	6.1	4.0	0.64	0.66
	0.8	M	92.7	93.7	1.2	2.5	6.0	3.8	0.55	0.58
		A	92.4	91.2	1.3	3.7	6.3	5.1	0.55	0.61
		R	92.5	93.0	1.4	3.0	6.1	4.0	0.59	0.57
	0.9	M	92.7	93.5	1.2	2.6	6.1	4.0	0.52	0.54
		A	92.8	92.7	1.2	2.2	6.1	4.8	0.52	0.53
		R	92.7	93.4	1.1	2.5	6.1	4.3	0.54	0.56
120	0.7	M	93.7	93.9	1.2	3.4	5.1	2.8	0.42	0.44
		A	93.4	91.6	1.4	3.6	5.2	4.8	0.42	0.43
		R	93.5	94.0	1.2	2.5	5.3	3.5	0.46	0.47
	0.8	M	93.7	94.2	1.3	3.0	5.1	2.9	0.39	0.41
		A	93.4	91.7	1.4	3.8	5.3	4.5	0.39	0.40
		R	93.6	94.2	1.2	3.0	5.1	3.0	0.42	0.43
	0.9	M	93.9	94.2	1.3	2.8	4.8	3.1	0.37	0.38
		A	93.7	93.7	1.4	2.9	4.9	3.5	0.37	0.38
		R	93.8	95.0	1.3	2.1	4.9	2.9	0.39	0.39

TABLE 2

Confidence interval coverage probability (CP), lower (L) and upper (U) tail error rates and average lengths (AL) for the distribution function $\theta = F(y) = 0.86$ with $p = 0.7$, 0.8, 0.9 and $n = 60$, 120: Imputation methods R and A; $R = 10,000$ simulations; $Y \sim \exp(1)$; NA = normal approximation, EL = empirical likelihood.

			CP(%)		L(%)		U(%)		AL	
n	p	IMP	NA	EL	NA	EL	NA	EL	NA	EL
60	0.7	R	90.9	95.1	7.9	2.3	1.2	2.6	0.22	0.27
		A	91.9	95.3	7.2	2.3	0.9	2.4	0.21	0.26
	0.8	R	92.8	94.7	6.0	2.4	1.2	2.8	0.20	0.25
		A	93.4	95.2	5.8	2.2	0.9	2.5	0.20	0.24
	0.9	R	92.3	94.7	6.7	2.8	1.0	2.5	0.19	0.22
		A	91.8	94.7	7.3	3.9	0.8	2.2	0.18	0.22
120	0.7	R	93.1	94.8	5.5	2.6	1.4	2.6	0.16	0.19
		A	93.0	95.0	5.9	2.7	1.1	2.3	0.15	0.18
	0.8	R	92.5	94.6	6.2	2.9	1.3	2.5	0.15	0.18
		A	93.2	94.9	5.8	2.8	1.1	2.3	0.14	0.17
	0.9	R	93.9	94.8	4.6	2.8	1.5	2.4	0.13	0.16
		A	94.0	94.9	4.6	2.6	1.4	2.5	0.13	0.16

TABLE 3

Confidence interval coverage probability (CP), lower (L) and upper (U) tail error rates and average lengths (AL) for the median $\theta_{\frac{1}{2}} = F^{-1}(\frac{1}{2})$ with $p = 0.7$, 0.8, 0.9 and $n = 60$, 120: Imputation methods R and A; $R = 10,000$ simulations; $Y \sim$ exp(1); NA = normal approximation, EL = empirical likelihood, W = Woodruff.

			CP(%)			L(%)			U(%)			AL		
n	p	IMP	NA	EL	W	NA	EL	W	NA	EL	W	NA	EL	W
60	0.7	R	87.3	95.0	95.5	4.0	2.5	2.4	8.7	2.4	2.2	0.65	0.68	0.70
		A	90.4	95.8	96.1	3.9	2.9	2.8	5.7	1.3	1.2	0.63	0.67	0.68
	0.8	R	87.3	95.3	95.6	4.1	2.3	2.3	8.6	2.4	2.2	0.60	0.63	0.64
		A	91.0	95.6	95.7	3.5	3.0	2.9	5.5	1.5	1.4	0.58	0.61	0.62
	0.9	R	88.5	95.4	95.4	3.2	2.3	2.3	8.3	2.3	2.2	0.54	0.58	0.58
		A	91.0	95.0	95.3	3.3	3.1	2.9	5.6	1.9	1.8	0.54	0.55	0.56
120	0.7	R	88.6	94.5	94.8	3.9	2.9	2.8	7.5	2.7	2.4	0.47	0.48	0.48
		A	91.5	95.2	95.5	3.4	2.9	2.8	5.1	1.9	1.7	0.45	0.46	0.47
	0.8	R	89.6	94.7	94.9	3.7	2.7	2.7	6.8	2.5	2.4	0.43	0.43	0.44
		A	92.1	95.6	95.7	3.3	2.8	2.8	4.7	1.6	1.5	0.41	0.42	0.43
	0.9	R	90.3	94.6	94.7	3.1	2.7	2.7	6.6	2.6	2.6	0.39	0.40	0.40
		A	91.7	95.0	95.1	3.2	3.1	3.1	5.1	2.0	1.9	0.38	0.39	0.39