Comparison of design-based sample mean estimate with an estimate under re-sampling-based multiple imputations

Recai Yucel

1 Introduction

This section introduces the general notation used throughout this report.

Let Y denote a binary random variable, and let the values of the Y in a random sample of n be denoted as $y = (y_1, y_2, \ldots, y_n)$. We assume that this random sample of n is obtained under a simple random sample without replacement (SRSWOR). Further we will work with the decomposition of y corresponding to the observed values and missing values: $y_{com} = (y_{obs}, y_{mis})$. Missingness indicator r_i will be used to in the following way:

$$r_i = \begin{cases} 1 & \text{if } y_i \text{ is missing,} \\ 0 & \text{if } y_i \text{ is observed,} \end{cases}$$

and $r = (r_1, r_2, \ldots, r_n)$. Methods dealing with missing data typically assume one of the following missingness mecahnisims:

$$\begin{aligned} &\text{MCAR: } P(r \mid y_{obs}, y_{mis}) = P(r) \\ &\text{MAR: } P(r \mid y_{obs}, y_{mis}) = P(r \mid y_{obs}) \\ &\text{MNAR: } P(r \mid y_{obs}, y_{mis}) = P(r \mid y_{obs}, y_{mis}) \end{aligned}$$

Throughout this report we will assume MCAR as the underlying mechanism for missingness. The general idea of multiple imputation is to replace missing values with m sets of plausible values. In a parametric multiple imputation, an imputation model (e.g. normal distribution) is used to draw these values, which is often called predictive distribution of missing values. To make a fair comparison of the estimation methods between design-based estimate by Stanek et al. , we will not assume any parametric structure on Y, but rather randomly sample from y_{obs}. The details are explained below.

2 Estimation routines

2.1 Stanek et al. estimate

The estimate of the population mean is proposed to be the weighted sum of three terms:

$$\hat{\mu}_0 = \frac{1}{N}[n\bar{Y} + (N-n)\hat{P}_1 + N\pi\hat{P}_2], \tag{1}$$

where

$\bar{Y} \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i$ sample mean (for missing values $Y_i = 0$, i.e. $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} r_i Y_i$)

$\hat{P}_1$: predictor of response for subject not selected ($\bar{Y}$)

$\hat{P}_2$: predictor of response for $N\pi$ subjects where the response is expected to be missing

π : is the estimate of the probability of responding

The estimate of the variance of this estimate is given by

$$\hat{V}(\hat{\mu}_0) = \frac{n_0}{nn_1}T^2 + \frac{N-n}{N}\frac{s_1^2}{n}, \tag{2}$$

where

$$\begin{aligned} T^2 &= \frac{1}{n_1}\sum_{i=1}^{n} r_i Y_i^2, \quad \text{where } n_1 = n_{obs},\, n_0 = n_{mis} \\ s^2 &= \text{sample variance based on } y_{obs}, \text{ assuming } y_{mis} = 0 \end{aligned}$$

2.2 Multiple imputation estimate

m sets of imputations are obtained by random draws from y_{obs} using SRSWOR. After obtaining m imputations of y_{mis}, we calculate the sample mean and estimate of its variance for each of the imputed dataset. These estimates are then combined using rules for scalar estimates by Rubin (1987). Note that these rules do not relate the procedure used in creating the imputations nor the missingness mechanism. It should be seen as a way to reflect the uncertainty due to imputation method into estimation. In standard notation, these rules are given below:

$$\begin{aligned}
\hat{Q} &= \text{complete-data point estimate} \\
\hat{U} &= \text{complete-data variance estimate} \\
\bar{Q} &= m^{(-1)} \sum_{t=1}^{m} \hat{Q}^{(t)} \\
B &= (m-1)^{-1} \sum_{t=1}^{m} (\hat{Q}^{(t)} - \bar{Q})^2 \\
&= \text{Between imputation variance} \\
\bar{U} &= m^{(-1)} \sum_{t=1}^{m} U^{(t)} \\
&= \text{Within imputation variance} \\
T &= \bar{U} + (1 + m^{-1})B \\
&= \text{Total variance}
\end{aligned}$$

Interval estimate is $\bar{Q} \pm t_\nu \sqrt{T}$, where

$$\nu = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2.$$

Degrees of freedom vary from $m-1$ to ∞, depending on relative sizes of $\bar{U}$ and $(1+m^{-1})B$. Relative increase in variance due to nonresponse is estimated by

$$r = \frac{(1+m^{-1})B}{\bar{U}},$$

and, fraction of missing information is estimated by $\frac{r+2/(\nu+3)}{r+1}$. It is often noted that this estimate can be noisy for small n

In our application, complete-data point estimate is given by $Q = \bar{y} = \sum_{i=1}^{n} y_i/n$ and complete-data variance estimate is given by $U = \hat{Var}(\bar{y}) = \frac{N-n}{N-1}\frac{s^2}{n}$, where t denotes the imputation number. Note that these are estimates under SRSWOR.

Question: Should one correct these estimates to reflect the fact that parts of data were imputed from y_{obs}?

3 Simulation study

3.1 Simulation conditions

This simulation study attempts to compare performances of the following estimators:

- design-based estimator by Stanek et al.
- Multiple imputation

These methods are explained in detail below in (2) and (4). Notation used is also explained below.

This simulation experiment assumes that the population consists of $N = 100$ binary values and simulations repeatedly draw sample of $n = 20$ via simple random sampling without replacement (SRSWOR). Let y_i denote the i^{th} value of the sampled unit, and let y denote the vector that consists of the y_i, $y = (y_1, \ldots, y_n)$.

Total number of repetition is 1000, and in each of the repetition we perform the following:

1. **Sampling** Select $n = 20$ from $N = 100$ using SRSWOR.
2. **Imposing missing values** Draw missingness indicator, $r_i \sim Bernoulli(0.6), i = 1, 2, \ldots, n$. Note that this indicator will be used to set the values of y_i to missing in

the following sense:

$$y_i = \begin{cases} 1 & \text{if } y_i \text{ is missing,} \\ 0 & \text{if } y_i \text{ is observed.} \end{cases}$$

Let y_{obs} and y_{mis} denote the partitions of y corresponding to observed and missing parts of y. Then $y_{obs} = y[r == 0]$.

3. **Drawing (re-sampling) imputations from y_{obs}.** In each cycle of the simulation, form multiple imputations, i.e. multiply re-sample $n^* = n - n_{obs}$ from y_{obs} using SRSWOR. This step consists of the following three steps:

   (a) Sample n_{mis} from n_{obs} using SRSWOR,

   (b) Calculate estimates of mean $(\bar{Y})$ and its variance $(\hat{Var}(\bar{y}))$ using standard SRSWOR formulas,

   (c) Repeat (a) and (b) 10 times, each time store the estimates,

   (d) Combine the 10 sets of mean estimate and its variance estimate.

3.2 Results and next steps

The results show consistency between two estimates with respect to evaluation criterion MSE. Note that the column BD (the estimates based on sample before deletion) represents the gold standard that the two approach try to capture. There is a gap between the MSEs of the two method and the MSE of the sample mean before deletion. It would be desirable to further understand whether this gap is important, and whether the estimates could be improved to close the gap. It is also important to further understand the differences in the variance estimates between design-based and MI methods. Surprisingly, the MI method resulted in estimates that were closer to estimates under BD.

Second step will be to look at the combined variance of the estimate under MI (column 2). This estimate is based on the following two quantities: Between imputation variance assessing the variability across the imputations $B = (m-1)^{-1}\sum_{t=1}^{m}(\hat{Q}^{(t)} - \bar{Q})^2 = (m -$

Table 1: Simulation results: Mean estimates followed by the $\sqrt{MSE}$, given in parantheses (BD: before deletion; MI: multiple imputation, Ed: Ed's method; all are averages across the simulations)

	Method		
	BD	MI	Ed
Scenario 1: μ=0.19, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.08816$			
	0.9015(0.0788)	0.1895(0.0993)	0.1895(0.0991)
Scenario 2: μ=0.35,$\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1072$			
	0.3489(0.0312)	0.3502(0.0389)	0.3504(0.0389)
Scenario 3: μ=0.57, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1113$			
	0.5692(0.0708)	0.5726 (0.0747)	0.5719(0.0745)
Scenario 4: μ=0.66, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1065$			
	0.6605(0.0301)	0.6589 (0.0384)	0.6591(0.0380)
Scenario 5: μ=0.72, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1009$			
	0.7227(0.0285)	0.7238 (0.0352)	0.7233(0.0354)
Scenario 6: μ= 0.8 , $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.0899$			
	0.7973 (0.0254)	0.7961 (0.0325)	0.7968 (0.0324)
Scenario 7: μ=0.91, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.0643$			
	0.9099(0.0178)	0.9104 (0.0227)	0.9106(0.0226)

Table 2: Simulation results: Variance estimates (BD: before deletion; MI: multiple imputation, Ed: Ed's method; all are averages across the simulations)

	Method		
	BD	MI	Ed
Scenario 1: μ=0.19, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.08816$			
	0.00774	0.00722	0.01015
Scenario 2: μ=0.35,$\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1072$			
	0.01144	0.01083	0.01685
Scenario 3: μ=0.57, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1113$			
	0.01235	0.01169	0.02262
Scenario 4: μ=0.66, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1065$			
	0.01133	0.01063	0.02369
Scenario 5: μ=0.72, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.1009$			
	0.01012	0.00954	0.02462
Scenario 6: μ= 0.8 , $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.0899$			
	0.00817	0.00767	0.02487
Scenario 7: μ=0.91, $\sigma_{\bar{Y}} = \sigma/\sqrt{n} = 0.0643$			
	0.00415	0.00389	0.02424

$1)^{-1}\sum_{t=1}^{m}(\bar{y}^{(t)}-\bar{\bar{y}})^2$, where $\bar{\bar{y}}$ is the average of the sample means across the imputations. The second quantity is the within imputation variance: $W = m^{(-1)}\sum_{t=1}^{m}U^{(t)}$. The total variance is calculated to be $\bar{U}+(1+m^{-1})B$ (Rubin, 1986). As discussed by Schenker and Rubin (1986), the factor $(1+m^{-1})$ reflects the extra variability due to imputations based on a finite number of imputations (small m). It will be important to derive the estimate of this variance from a pure finite sampling point in which several processes needed to be taken into account: sampling, missingness mechanism and imputation. This step is also important in extending the re-sampling-based multiple imputation inference under other sampling schemes such as clustered or stratified designs.

Final step pertains to extending the design-based and MI approaches to multivariate settings. Creating imputations by resampling from y_{obs} will be somewhat cumbersome under the arbitrary missingness, and developing (or using previous methods) sound algorithmical rules (such as matching to propensity scores) would be potential contributions.

References

Rubin, D.B. (1986), Multiple imputation for Survey Nonresponse, New York, John Wiley.

Rubin, D.B. and Schenker, N. (1986), "Multiple imputation for interval estimate from simple random samples with igorable nonresponse“, Journal of the American Statistical Association, Vol. 81, No. 394, 366–374.