VARIANCE OF THE ESTIMATED MEAN FOR SEVERAL IMPUTATION PROCEDURES

Lawrence R. Ernst, U.S. Bureau of the Census

1. INTRODUCTION

Many imputation procedures for missing observations for a variate proceed by dividing the sample into adjustment cells based on ancillary variables assumed known for all sample cases, and then substituting for each missing observation in an adjustment cell a value which is a linear combination of the observed values in that adjustment cell. Since the variance of the imputation mean for the entire sample is dependent on the variances of the imputation means for the individual adjustment cells, we will concern ourselves with determining the variance of the imputation mean for an adjustment cell. Consequently the entire sample will be assumed to consist of a single adjustment cell.

We will consider six such imputation procedures, which are defined in Section 2: mean of the observations (MO), running mean (RM), previous observation (PO), nearest observation (NO), systematic imputation (SI), and random imputation (RI). In the last four of these procedures each imputed value will always be some observed value; such procedures will be referred to as hot deck procedures.

Surprisingly, relatively little has been done for some of these procedures in developing variance formulas and comparing the variances. For example, for the RM and NO procedures this author is aware of no previous work of this type. For PO, which is probably the most commonly used of the hot deck procedures, the only previously derived formulas appeared in Bailar and Bailar (1978). In that work the variances were conditional on the number of missing values and the combinational approach taken resulted in rather difficult proofs. For the other three procedures variance formulas are more easily obtainable, but are usually not presented in the form given here.

In this paper variance formulas are derived for all six procedures. Following the approach taken in Bailar and Bailar (1978) this is done in two cases, uncorrelated (UC) and serially correlated (SC), which differ in their assumptions on the correlation structure of the sample. For PO and NO both exact and asymptotic formulas are derived; although for NO the exact formulas are presented only in the Appendix due to their complexity. (Copies of the Appendix are available from the author.) For the other four procedures exact variance expressions are not obtainable and, therefore, only asymptotic formulas are presented.

Comparisons between the asymptotic variances of these six procedures are then made. Among the results is that if $AV(\bar{x}_\alpha)$ denotes the asymptotic variance with respect to procedure α, and the probability of an observation being missing does not exceed 1/3, then

$$AV(\bar{x}_{MO}) \leq AV(\bar{x}_{RM}) \leq AV(\bar{x}_{SI})$$

$$\leq AV(\bar{x}_{NO}) \leq AV(\bar{x}_{RI}) \leq AV(\bar{x}_{PO})$$

in the UC case, and also in the SC case when the correlation coefficient for adjacent units, ρ, is sufficiently small (Theorem 13); while

$$AV(\bar{x}_{NO}) \leq AV(\bar{x}_{PO}) \leq AV(\bar{x}_{MO})$$

$$\leq AV(\bar{x}_{RI}) \leq AV(\bar{x}_{RM}) \leq AV(\bar{x}_{SI})$$

in the SC case when ρ is sufficiently large (Theorem 14). Different relations hold if the probability of an observation being missing is greater than 1/3, as detailed in these two theorems.

2. NOTATION, TERMINOLOGY AND ASSUMPTIONS

The initial assumptions on the sampling procedure are that it yields a self-weighting sample of size n drawn from a population with mean μ and variance σ^2. Let m denote the number of missing values in a given sample, and $q = E(m)/n < 1$.

For $i = 1, \ldots, n$ let x_i denote the variate value for the i-th sample unit, whether observed or not, and let $w_i = 1$ if x_i is observed, $w_i = 0$ if x_i is missing. It is not assumed that the sample units are in random order. However, it is assumed that each x_i has the same distribution as the population, that x_i and w_i are independent for all i, and that for a fixed m all possible arrangements of the missing values are equally likely. In addition the following alternate set of assumptions are made on the structure of $\mathrm{cov}(x_i,x_j)$ for $i \neq j$ to distinguish our two cases:

Uncorrelated (UC). $\mathrm{cov}(x_i,x_j) = 0$. Observe that simple random sampling with replacement will have this covariance structure and also satisfy all previous assumptions in this section.

Serially Correlated (SC). $\mathrm{cov}(x_i,x_j) = \rho^{|i-j|}\sigma^2$, where $0 \leq \rho < 1$. Note that the SC case reduces to the UC case for $\rho = 0$.

For the PO and RM procedures we employ an additional value x_0, known as the cold deck value. This value is chosen randomly from the same population as the sample, but independently of it.

For any real number x, let $\lfloor x \rfloor$ denote the greatest integer not exceeding x; then let $r = 1/(1-q) - \lfloor 1/(1-q) \rfloor$. These definitions will be used for the SI procedure.

We now proceed to define the six imputation procedures previously referred to. For the definitions of the PO, NO, SI, and RI procedures recall that the sample units have been sequenced.

Mean of the Observations (MO). For each unobserved value, the imputed value is the same, namely the mean of all observed values. Note that in terms of mean and variance this procedure is identical to ignoring the missing values.

Previous Observation (PO). The immediately preceding observed value in the sequence is imputed for each missing value. If there are no preceding observed values, then x_0 is imputed.

Nearest Observation (NO). An observed value in the sequence nearest to the missing value is imputed. If there are two such observed values then one of them is imputed randomly.

Systematic Imputation (SI). The observed values are used sequentially for imputation. That is if $m < n - m$ then the i-th observed value will be imputed for the i-th unobserved value. If $m > n - m$ start over with the first observed value after each use of the (n-m)-th observed value. Thus, in general the $(i - (n-m)\lfloor (i-1)/(n-m) \rfloor)$-th observed value is imputed for the i-th unobserved value.

Random Imputation (RI). An observed value is chosen randomly with replacement to substitute for each missing value.

Running Mean (RM). The mean of all preceding observed values in the sequence is imputed. If there are no preceding observed values then x_0 is imputed.

Note that the imputation means for the MO, NO, SI, and RI procedures are not defined if $w_i = 0$ for all i. To avoid this problem we consider these four imputation means to be condititional on $w_i = 1$ for some i.

We also observe that with the assumptions previously given on the x_i's and w_i's each of the six procedures yields an unbiased estimator of μ.

If α is any of the six procedures then $\bar{x}_\alpha$ denotes the imputation mean with respect to procedure α, while $V(\bar{x}_\alpha)$ denotes the variance of $\bar{x}_\alpha$ for either the UC or SC cases. The asymptotic variance of $\bar{x}_\alpha$, denoted $AV(\bar{x}_\alpha)$, is given by

$$AV(\bar{x}_\alpha) = \frac{\left[\lim_{n\to\infty} nV(\bar{x}_\alpha)\right]}{n}.$$

3. SUMMARY OF RESULTS

In this section variance and/or asymptotic variance formulas are presented for the six procedures in the two cases. All proofs are deferred until the Appendix.

3.1 Variance and Asymptotic Variance Formulas in UC Case

Theorem 1: $$AV(\bar{x}_{MO}) = \frac{\sigma^2}{n(1-q)}.$$

Theorem 3: $$V(\bar{x}_{PO}) = \frac{\sigma^2}{n}\left[\frac{1+q}{1-q} + \frac{-2q+2q^{n+1}}{n(1-q)^2}\right],$$

and hence $$AV(\bar{x}_{PO}) = \frac{\sigma^2}{n}\left(\frac{1+q}{1-q}\right).$$

Theorem 5: $$AV(\bar{x}_{NO}) = \frac{\sigma^2}{n}\left[\frac{2+4q-q^2+q^3}{2(1-q^2)}\right].$$

Theorem 7: $$AV(\bar{x}_{SI}) = \frac{\sigma^2}{n}\left[\frac{1+(1-q)^2r(1-r)}{1-q}\right].$$

Theorem 9: $$AV(\bar{x}_{RI}) = \frac{\sigma^2}{n}\left(\frac{1+q-q^2}{1-q}\right).$$

Theorem 11: $$AV(\bar{x}_{RM}) = \frac{\sigma^2}{n}\left(\frac{1+q^2}{1-q}\right).$$

3.2 Variance and Asymptotic Variance Formulas in SC Case

Theorem 2: $$AV(\bar{x}_{MO}) = \frac{\sigma^2}{n}\left(\frac{1}{1-q} + \frac{2\rho}{1-\rho}\right).$$

Theorem 4:

$$V(\bar{x}_{PO}) = \frac{\sigma^2}{n}\left(\left[\frac{1+q}{1-q} + \frac{2\rho}{1-\rho} - \frac{2q\rho}{1-q\rho}\right]\right.$$

$$+ \frac{2}{n}\left[\frac{-q+q^{n+1}}{(1-q)^2} + \frac{-\rho+\rho^{n+1}}{(1-\rho)^2} + \frac{q\rho-(q\rho)^{n+1}}{(1-q\rho)^2}\right]$$

$$+ \frac{2}{n(1-q)}\left[\frac{(q+q^{n+1})(-\rho+\rho^n)}{1-\rho}\right.$$

$$\left.\left. + \frac{q^2\rho-q^{n+1}\rho^n}{1-q\rho} + \frac{q^{n+1}\rho-q^2\rho^n}{q-\rho}\right]\right),$$

and hence $$AV(\bar{x}_{PO}) = \frac{\sigma^2}{n}\left(\frac{1+q}{1-q} + \frac{2\rho}{1-\rho} - \frac{2q\rho}{1-q\rho}\right).$$

Theorem 6:

$$AV(\bar{x}_{NO}) = \frac{\sigma^2}{n}\left[\frac{2+4q-q^2+q^3}{2(1-q^2)} + \frac{2\rho}{1-\rho}\right.$$

$$\left. + \frac{(-4q+q^2)\rho+3q^2\rho^2+(3q^2-2q^3+2q^4)\rho^3+(-q^3-q^4-q^5)\rho^4}{2(1-q\rho)^3(1+q\rho)}\right].$$

Theorem 8:

$$AV(\bar{x}_{SI}) = \frac{\sigma^2}{n}\left[1+(1-q)^2r(1-r)\right]\left(\frac{1}{1-q} + \frac{2\rho}{1-\rho}\right)$$

Theorem 10: $$AV(\bar{x}_{RI}) = \frac{\sigma^2}{n}\left(\frac{1+q-q^2}{1-q} + \frac{2\rho}{1-\rho}\right).$$

Theorem 12: $$AV(\bar{x}_{RM}) = \frac{\sigma^2}{n}(1+q^2)\left(\frac{1}{1-q} + \frac{2\rho}{1-\rho}\right).$$

4. COMPARISONS

We state the results of the comparisons between the asymptotic variances, and present a table illustrating these comparisons. Again all proofs are given in the Appendix.

Theorem 13: In the UC case and also in the SC case with sufficiently small ρ the following relations hold:

$$AV(\bar{x}_{MO}) \le AV(\bar{x}_{RM}) \le AV(\bar{x}_{SI}) \le AV(\bar{x}_{NO}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{PO}) \text{ if } q \le 1/3;$$

$$AV(\bar{x}_{MO}) \le AV(\bar{x}_{SI}) \le AV(\bar{x}_{RM}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{NO}) \le AV(\bar{x}_{PO}) \text{ if } 1/3 \le q \le 1/2;$$

$$AV(\bar{x}_{MO}) \le AV(\bar{x}_{SI}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{RM}) \le AV(\bar{x}_{NO}) \le AV(\bar{x}_{PO}) \text{ if } 1/2 \le q \le (-3+\sqrt{17})/2;$$

$$AV(\bar{x}_{MO}) \le AV(\bar{x}_{SI}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{NO}) \le AV(\bar{x}_{RM}) \le AV(\bar{x}_{PO}) \text{ if } q \ge (-3+\sqrt{17})/2.$$

Theorem 14: In the SC case with sufficiently large ρ the following relations hold:

$$AV(\bar{x}_{NO}) \le AV(\bar{x}_{PO}) \le AV(\bar{x}_{MO}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{RM}) \le AV(\bar{x}_{SI}) \text{ if } q \le 1/3;$$

$$AV(\bar{x}_{NO}) \le AV(\bar{x}_{PO}) \le AV(\bar{x}_{MO}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{SI}) \le AV(\bar{x}_{RM}) \text{ if } q \ge 1/3 \text{ and } 1/(1-q) \text{ is not an integer;}$$

$$AV(\bar{x}_{NO}) \le AV(\bar{x}_{PO}) \le AV(\bar{x}_{MO}) = AV(\bar{x}_{SI}) \le AV(\bar{x}_{RI}) \le AV(\bar{x}_{RM}) \text{ if } 1/(1-q) \text{ is an integer.}$$

Although not explicitly stated in Theorems 13 and 14, the proofs in the Appendix actually yield somewhat stronger relations in the SC case for any pair of procedures α, β and any fixed q as follows: If $AV(\bar{x}_\alpha) \le AV(\bar{x}_\beta)$ for ρ sufficiently small and for ρ sufficiently large, then $AV(\bar{x}_\alpha) \le AV(\bar{x}_\beta)$ for all ρ. On the other hand, if $AV(\bar{x}_\alpha) \le AV(\bar{x}_\beta)$ for ρ sufficiently small and $AV(\bar{x}_\beta) \le AV(\bar{x}_\alpha)$ for ρ sufficiently large, then there exists ρ_o such that $AV(\bar{x}_\alpha) \le AV(\bar{x}_\beta)$ for $\rho \le \rho_o$ and $AV(\bar{x}_\beta) \le AV(\bar{x}_\alpha)$ for $\rho \ge \rho_o$.

The following table illustrates the results of the previous two theorems. It gives the asymptotic efficiencies with respect to $\bar{x}_{MO}$ of the other five estimators in the SC case for $\rho = 0, .1, \ldots, .9$ ($\rho = 0$ is the UC case of course), and $q = .1, .2, \ldots, .9$. The reason that asymptotic efficiencies with respect to MO were computed, is that in terms of variance the MO procedure is, as previously noted, identical to ignoring missing observations.

5. DISCUSSION OF RESULTS

We briefly discuss each of the six procedures in terms of the results of Theorems 13 and 14, and then state our conclusions. Some of the comparisons will be between the four hot deck procedures only. This is because these four imputation procedures, unlike the other two, preserve the marginal distribution of x_i for each

ASYMPTOTIC EFFICIENCY OF ESTIMATORS WITH RESPECT TO $\bar{x}_{MO}$

Estimator	ρ	q = .1	.2	.3	.4	.5	.6	.7	.8	.9
$\bar{x}_{RM}$	All	.9901	.9615	.9174	.8621	.8000	.7353	.6711	.6098	.5525
$\bar{x}_{SI}$	All	.9259	.8929	.8929	.9259	1.0000	.9615	.9804	1.0000	1.0000
$\bar{x}_{NO}$	.0	.9201	.8671	.8288	.7991	.7742	.7519	.7307	.7098	.6885
	.1	.9456	.9060	.8742	.8461	.8191	.7918	.7632	.7328	.7005
	.2	.9661	.9398	.9163	.8925	.8660	.8356	.8003	.7600	.7152
	.3	.9820	.9679	.9538	.9366	.9136	.8828	.8427	.7928	.7335
	.4	.9938	.9899	.9852	.9763	.9599	.9325	.8908	.8325	.7570
	.5	1.0019	1.0058	1.0092	1.0092	1.0019	.9822	.9440	.8808	.7881
	.6	1.0066	1.0154	1.0248	1.0327	1.0356	1.0277	1.0001	.9394	.8307
	.7	1.0084	1.0191	1.0315	1.0445	1.0562	1.0621	1.0525	1.0076	.8918
	.8	1.0077	1.0173	1.0291	1.0431	1.0590	1.0754	1.0870	1.0756	.9819
	.9	1.0047	1.0107	1.0183	1.0279	1.0403	1.0568	1.0786	1.1040	1.0968
$\bar{x}_{RI}$	.0	.9174	.8621	.8264	.8065	.8000	.8065	.8264	.8621	.9174
	.1	.9302	.8804	.8462	.8252	.8163	.8194	.8355	.8672	.9191
	.2	.9416	.8974	.8654	.8442	.8333	.8333	.8456	.8730	.9211
	.3	.9517	.9133	.8840	.8632	.8511	.8484	.8569	.8798	.9235
	.4	.9607	.9281	.9020	.8824	.8696	.8647	.8696	.8879	.9264
	.5	.9689	.9420	.9195	.9016	.8889	.8824	.8840	.8974	.9302
	.6	.9763	.9551	.9366	.9211	.9091	.9016	.9005	.9091	.9353
	.7	.9830	.9673	.9531	.9406	.9302	.9227	.9195	.9236	.9422
	.8	.9891	.9788	.9692	.9603	.9524	.9459	.9418	.9420	.9524
	.9	.9948	.9897	.9848	.9801	.9756	.9716	.9682	.9664	.9689
$\bar{x}_{PO}$	.0	.9091	.8333	.7692	.7143	.6667	.6250	.5882	.5556	.5263
	.1	.9362	.8756	.8182	.7640	.7129	.6648	.6196	.5771	.5373
	.2	.9582	.9130	.8650	.8147	.7627	.7097	.6563	.6031	.5509
	.3	.9756	.9451	.9083	.8650	.8153	.7599	.6993	.6349	.5679
	.4	.9888	.9714	.9466	.9130	.8696	.8153	.7500	.6743	.5900
	.5	.9981	.9915	.9784	.9565	.9231	.8750	.8093	.7241	.6197
	.6	1.0040	1.0054	1.0024	.9925	.9722	.9362	.8774	.7879	.6615
	.7	1.0069	1.0129	1.0172	1.0179	1.0117	.9928	.9510	.8692	.7239
	.8	1.0069	1.0144	1.0220	1.0292	1.0345	1.0341	1.0191	.9669	.8235
	.9	1.0046	1.0099	1.0163	1.0239	1.0329	1.0431	1.0529	1.0523	.9834

i. For this reason alone one might limit the choice among the six procedures to these four.

MO. This procedure has the smallest asymptotic variance of all six under the conditions of Theorem 13. However, it has a larger asymptotic variance than NO and PO under the conditions of Theorem 14 for all q.

PO. It has the largest asymptotic variance of all six procedures under the conditions of Theorem 13, but its asymptotic variance is smaller than all but NO under the conditions of Theorem 14. Interestingly, despite its wide usage, PO never has the smallest asymptotic variance among the four hot deck procedures.

NO. Although not distinguished under the conditions of Theorem 13, it has the smallest asymptotic variance of the six procedures under the conditions of Theorem 14 for all q.

SI. This procedure has been used with such surveys as the March Current Population Survey of the Bureau of the Census as a lower variance alternative to PO. Indeed, under the conditions of Theorem 13 it has the smallest asymptotic variance of the four hot deck procedures. However, under the conditions of Theorem 14 its asymptotic variance is relatively large.

RI. Although it never has the largest asymptotic variance among the six procedures, it also never has the smallest, either among all six procedures or the four hot deck procedures.

RM. Under the conditions of Theorem 13 it has a smaller asymptotic variance than all but MO if $q < 1/3$, but does not perform as well for larger q. It also has a relatively large asymptotic variance under the conditions of Theorem 14 for all q.

On the basis of smallest asymptotic variance MO is best under the conditions of Theorem 13 among all six procedures and SI among the four hot deck procedures. Under the conditions of Theorem 14, NO is the best both among all six procedures and the four hot deck procedures.

Other considerations, however, may enter into the choice of procedures. For example, PO and RM are the easiest to program.

Furthermore, different assumptions than those made in this paper will often lead to different conclusions on choice of imputation procedure. In particular these six procedures will, in general, no longer yield unbiased estimators of μ if one drops either the assumption that x_i and w_i are independent for all i, or the assumption that $E(x_i) = \mu$ for all i. In fact, the bias properties of these procedures may then become more significant than the variance properties, particularly for large samples. Bailar and Bailar (1979) have compared the biases of the MO and PO procedures under various assumptions, but much remains to be done. Unfortunately, it appears that most reasonable sets of assumptions which produce biased estimators also result in variances that are not mathematically tractable.

REFERENCES

Bailar, Barbara A., and Bailar, John C. III, (1979), "Comparison of the Biases of the 'Hot-Deck' Imputation Procedure with an 'Equal Weights' Imputation Procedure," Symposium on Incomplete Data: Preliminary Proceedings, Panel on Incomplete Data, Committee on National Statistics, National Research Council.

Bailar, John C. III, and Bailar, Barbara A., (1978), "Comparison of Two Procedures for Imputing Missing Survey Values," Proceedings of the Section on Survey Research Methods, American Statistical Association, 462-467.

Chapman, David W., (1976), "A Survey of Nonresponse Imputation Procedures," Proceedings of the Social Statistics Section, American Statistical Association, 245-251.

Ford, Barry L., (to appear), "An Overview of Hot Deck Procedures," Chap. 4, Incomplete Data: Theory of Current Practice, Panel on Incomplete Data, Committee on National Statistics, National Research Council.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G., (1953), Sample Survey Methods and Theory, Vol. II, New York: John Wiley and Sons.