COMPARISON OF HOT-DECK VARIATIONS IN INPUTING MISSING VALUES

Javaid Kaiser, El Paso Community College

Hot deck method is one of the most frequently used techniques for imputation of missing values. Census Bureaus of the United States and Canada, the Internal Revenue Service, and Social Security Administration are heavy users of this method. The relative efficiency of hot-deck variations in producing missing value estimates and the impact of imputing these estimates on the covariance structure of the data matrix are not completely known.

The previous work to explore hot-deck properties may be classified into: (1) analytical studies (Bailar & Bailar, 1978; Ernst, 1980; Kalton & Kasprzyk, 1983; Kalton & Kish, 1981; Proctor, 1978), (2) studies that used special data sets (Champney & Bell, 1982; Cox & Folsom, 1978; Ernst, 1978; Huddleston & Hocking, 1978), and (3) studies that have used simulation to generate data (Ford, Kleweno, & Tortora, 1982). Oh and Scheuren (1980) have provided the history of the development of hot-deck method and have compared it with other imputation methods.

The studies cited in the literature had two major limitations. First, is the use of univariate approach when the problem is multivariate in nature. Second, is the use of special data sets that restricted the generalizability of results across other applications. Simulation of data, when used, was very limited in scale. Analytical studies focused on a small aspect of the problem and failed to relate proposed solutions and the assumptions that were made in these studies to other areas of the problem. Most of these studies also lacked in providing empirical support to their findings. The present study was, therefore, designed to investigate the relative efficiency of hot-deck variations and cell mean method in terms of (1) estimating complete sample means, (2) computing quality estimates of missing values, (3) retaining complete sample covariance structure in imputed samples, and (4) keeping imputed samples representative of population covariance structure. Multivariate approach and simulated data sets were used to increase the generalizability of results.

The hot-deck variations investigated in this study were hot-deck sequential, hot-deck random, and hot-deck distance. The variations differed from one another in terms of selecting a donor record. Hot-deck sequential used immediately preceding complete record as donor. Hot-deck random selected the donor record at random from the complete records present in the stratum. In hot-deck distance, the donor was the nearest complete record and was not necessarily the immediately preceding complete record. When two records were equidistant, the mean of the two observed values was used as an estimate of missing value. The cell mean method imputed the cell mean observed on a vector as an estimate of missing values that occurred on that vector. All these methods, required at least one stratification variable to create cells.

Method

Three hot-deck variations were compared with cell mean method in a simulation environment with 3 X 3 X 4 factorial design. The factors studied were sample size: n = 30, 60, and 120; the proportion of incomplete records in a sample: p = .1, .2, and .3, and the number of missing values in a record: m = 1, 2, 3, and 4. The design matrix had a total of 36 cells and each cell was replicated 500 times.

An 11 X 11 correlation matrix was selected from the literature and was used as population correlation matrix for the purpose of generating data. The matrix had intercorrelation in the range of .19 to .47. The second and third vector of this matrix had a correlation of .25 and median correlation of these vectors with the next eight vectors was .27. Data matricesofmultivariate normal deviates of size n X 11 were generated from this population and were regarded as multivariate normal random samples in standard score form. A variance-covariance matrix was computed for each data matrix generated and was compared with the population variance-covariance matrix as described by Anderson (1958). The data matrices that had the population covariance structure ($P > .05$) were retained for this study. Five hundred data matrices were used for each cell of the design matrix.

The first variable on the data matrices was used as work vector for housekeeping activities. The next two variables were recoded from an interval scale to the nominal scale for use as stratification variables. The cut-off points were established arbitrarily such that $c = 1$ when $i \leq -1.0$, $c = 2$ when $-1.0 < i < 1.0$, and $c = 3$ when $i \geq 1.0$, where c and i represent nominal category created and the initial value observed on the vector, respectively. The combination of c values on the two stratification variables created a total of 9 strata.

Every record in the data matrix was assigned a stratum based on the values on stratification variables. The remaining n X 8 submatrix was used for imputation purposes.

Random variables were generated from a uniform distribution (0,1) first, to randomly select n.p records and then to randomly select n.m vectors on each of these selected records to represent missing values. The initial observed values on these data points were replaced by missing value code. Missing values were coded only on the last eight vectors of data matrices. The values of n, p, and m were determined from the cell specifications of the design matrix.

The variance-covariance matrix and means were computed on the complete sample before creating missing values and on the imputed sample after imputing missing data. The discrepancy in means between the complete sample and imputed sample as well as the variance of this discrepancy was computed on all replications to study distributional properties. The root-mean-square deviation of off-diagonal elements of variance covariance matrices representing complete and imputed samples was computed to determine the extent covariance structure of complete sample is retained in imputed sample. This statistic commonly known as $\underline{D}$ was first proposed by Timm (1970) and was later modified by Gleason & Staelin (1975). The statistic $\underline{Q}$ representing root-mean-square standardized residual between the actually observed and imputed values was computed to determine the quality of missing value estimates produced by an imputation method (Gleason et al. 1975). The variance-covariance matrix of imputed sample was also compared with the population variance-covariance matrix as described by Anderson (1958) to determine if the imputed sample retained population covariance structure.

Imputation methods were applied one at a time on data matrices after creating missing values. Donor records were not used more than once for hot-deck variations. The cell mean method used three subgroups which were created by recoding of the second vector of data matrices. The statistic collected on each method was averaged over 500 replications. The procedure, described above, was repeated for all cells of the design matrix.

Results

The discrepancy in means between complete and imputed samples was plotted for all experimental conditions and is given in Figure 1. The results revealed no substantial differences in hot-deck variations and cell mean method. The discrepancy in means for all these methods under all experimental conditions was in the range of $\pm$ .008. It was also observed that the range of average discrepancy in means increased with the increase in p, m, or their combination, and decreased with the increase in sample size. There was no systematic trend in the direction of the discrepancy in means as p, m, or n increased.

Though all the imputation methods performed well, their relative efficiency in imputing missing values became more visible when 20% or more records were incomplete and each one of them had 25% or more missing values. When the proportion of missing values was 50% with 20% or more incomplete records in the sample, the differences in the performance of imputation methods were the greatest. This was found true at all levels of n.

The relative rank of imputation methods varied under experimental conditions but no systematic trend was observed. In general, cell mean method performed better in most of the experimental conditions created by p and m at all levels of n. Hot-deck random method performed poorly at n = 30, but improved its efficiency as the sample size increased. Hot-deck sequential method performed generally poor at n > 30. Hot-deck distance method performed generally better than hot-deck sequential and almost equally as well as cell mean method.

Standard deviation of the discrepancy in means between complete and imputed sample increased with the increase in p or m and decreased with the increase in sample size. The cell mean method produced the lowest standard deviation. Hot-deck distance method was the second in rank. Hot-deck sequential and hot-deck random methods performed almost alike and were ranked the lowest in terms of reducing variance.

The analysis on statistic $\underline{D}$ revealed that the efficiency of all imputation methods in retaining covariance structure of complete samples in imputed samples decreased with the increase in p or m, and increased with the increase in sample size. The results were plotted for all experimental conditions and are displayed in Figure 2. There were no considerable differences among imputation methods in terms of their efficiency. However, the cell mean method performed the best followed by hot-deck distance method. The efficiency of the cell mean method dropped to the lowest of all methods when $p \geq .2$ and $m > 2$ at n = 120.

The statistic $\underline{Q}$ revealed that the quality of missing value estimates deteriorated slightly with the increase in p and m at n = 30, and that this trend was neutralized in large samples where the level of efficiency stayed the

same at all levels of m. It was observed that substitution by cell mean method produced the best estimates of missing values at all levels of n, p, and m. Hot-deck distance method was found as the second best imputation technique. Hot-deck sequential and hot-deck random were rated the lowest at all levels of p, m, and n. The plotted data on statistic Q is displayed in Figure 3.

Figure 4 describes the impact of imputation on the covariance structure of imputed samples. The results revealed that the increase in p, m, or their combination increased the percent of imputed samples that had significantly different covariance structure ($P \leq .05$) than that of the population from which they were drawn. The increase in sample size, however, reduced this adverse effect. The frequency of these affected samples increased sharply when the proportion of incomplete records was more than 10% and the missing values per record exceeded 25%.

In general, hot-deck distance method performed the best in retaining population covariance structure in imputed samples ($P > .05$). The cell mean method was found the second best. It produced a relatively large number of samples with deviant covariance structure ($P \leq .05$) than hot-deck distance method when samples contained 30% incomplete records and each record had 50% missing values. However, in small samples ($n \leq 60$) the cell mean method performed at least as good as hot-deck distance method. Hot-deck random method ranked the lowest. The range of data matrices with deviant covariance structure ($P \leq .05$) was 4.4% to 36.2% across all experimental conditions.

Discussion

The data analysis revealed that the average discrepancy between means of complete and imputed samples was almost zero ($\pm$ .008) for all imputation methods under all experimental conditions. It confirmed the finding of Champney et al. (1982) and Kalton et al. (1983) that all imputation methods are robust in estimating sample means. Since the discrepancy of $\pm$.008 in means is not more than what one may get as sampling error, the estimates of population means from imputed samples may not be very different from estimates obtained from complete sample irrespective of the imputation method used. This result confirmed the findings of Ernst (1980) and Hinkins (1983) that imputation methods yield unbiased estimates of population means.

Although imputing central value distorts distribution (Kalton et al., 1981; Kalton et al., 1983; Proctor, 1978), the cell mean method produced lower variance than hot deck variations. Champney et al. (1982) recorded the same result. The finding that hot-deck variations yield higher variance was also observed by Cox et al. (1978). Hinkins (1983) suggested that increased variance may be minimized with the use of hot-deck error term. The increase in variance with the increase in p and m indicated that estimates of population means may be poorer if the number of incomplete records, number of missing values in a record, or both, increases.

The results obtained from statistics D and Q confirmed that the cell mean method performed better than hot deck variations in producing quality estimates for missing values and in retaining complete sample covariance structure in imputed samples. No literature support was available on these findings as previous studies did not compare imputation methods in this context.

In ranking imputation methods, the cell mean method may be considered the best in estimating population and complete sample means, producing quality estimates of missing values, and in retaining complete sample covariance structure in imputed samples. Hot-deck distance method, on the other hand, seemed the best in retaining population covariance structure in imputed samples. Hot-deck sequential and hot-deck random methods were ranked the lowest. Although imputation methods have been ranked in earlier studies, the results of this study may not be compared with those as this study investigated different imputation methods and compared them in different contexts.

There were two findings of this study that have stayed consistent with previous literature. First, that hot-deck sequential method is comparatively an inferior technique, and, second, that hot-deck distance method is superior to hot-deck sequential and perhaps the best among hot-deck variations. Bailar et al. (1978), Cox et al. (1978), Ernst (1978 & 1980), Proctor (1978), and Schieber (1978), supported these findings.

The poor performance of hot-deck sequential method may be attributed to random samples and low serial correlation. Bailar et al. (1978) found that hot-deck sequential is unsuitable for random samples which was the case in this study. Moreover, as suggested by Bailar et al. (1978), no attempt was made to introduce serial correlation except that which was already present. It was also felt that using a large number of stratification variables having high correlation with imputable vectors may improve the overall performance of hot-deck variations.

Conclusion

The results revealed that all imputation methods are robust in estimating population or complete sample means. The cell mean method was found the best imputation technique in estimating missing values and in retaining complete sample covariance structure in imputed samples. Hot-deck distance method was considered the best in retaining population covariance structure in imputed samples. Hot-deck sequential was considered inferior to hot-deck distance and cell mean methods. Hot-deck random was ranked the lowest. The ranking of imputation methods was fairly consistent across all experimental conditions. There were no considerable differences among hot-deck variations and cell mean method in terms of the discrepancy in means between the complete and imputed samples, the variance of this discrepancy, and in retaining complete sample covariance structure in imputed samples. There were, however, substantial differences among imputation techniques in retaining population covariance structure in imputed samples and in producing estimates for missing values.

The data indicated that the increase in missing values per record, the proportion of incomplete records per sample, or their combination, severely affected the quality of missing value estimates, the retention of population and complete sample covariance structure in imputed samples, the magnitude of the discrepancy in means between complete and imputed samples, and the variance of this discrepancy. The increase in sample size minimized these adverse effects to some extent.

Imputation of data was found useful if the purpose is to estimate missing values or to estimate complete sample or population means. Imputing missing values with estimates did not seem a good choice if the purpose is to complete the data matrix for statistical analysis and to test hypotheses.

References

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. N.Y.: John Wiley & Sons, Inc.

Bailar, J. C., & Bailar, B. A. (1978). Comparison of two procedures for imputing missing survey values. Proceedings of the Section on Survey Research Methods, ASA (462-467).

Champney, T. F. & Bell, R. (1982). Imputation of Income: A procedural comparison. Proceedings of the Section on Survey Research Methods, ASA (431-436).

Cox, B. G. & Folsom, R. E. (1978). An empirical investigation of alternate item nonresponse adjustments. Proceedings of the Section on Survey Research Methods, ASA.

Ernst, L. R. (1978). Weighting to adjust for partial non-response. Proceedings of the Section on Survey Research Methods, ASA (468-472).

Ernst, L. R. (1980). Variance of the estimated mean for several imputation procedures. Proceedings of the Section on Survey Research Methods, ASA (716-720).

Ford, B. L., Kleweno, D. G., & Tortora, R. D. (1982). The effects of procedures which impute for missing item: A simulation study using an argicultural survey. Proceedings of the Section on Survey Research Methods, ASA (251-257).

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40:2 229-252.

Hinkins, Susan M. (1983) Imputation of missing items on corporate balance sheets. Proceedings of the Section on Survey Research Methods, ASA (254-259).

Huddleston, H. F. & Hocking, R. R. (1978). Imputation in agricultural survey. Proceedings of the Section on Survey Research Methods, ASA.

Kalton, G. & Kasprzyk, D. (1983). Imputing for missing survey responses. Proceedings of the Section on Survey Research Methods, ASA (22-31).

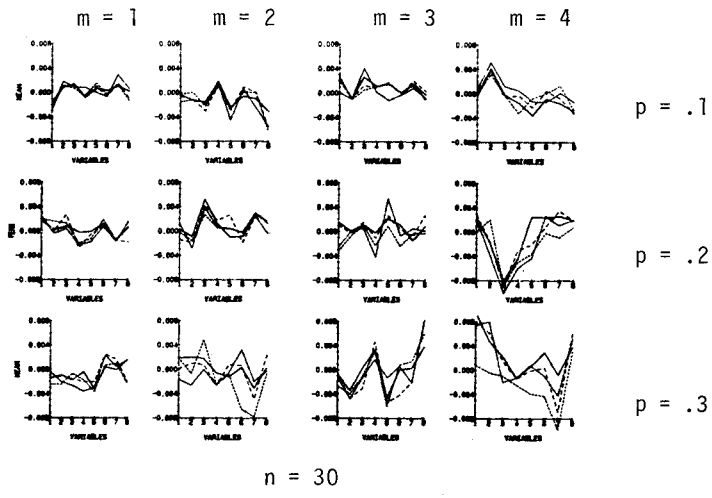
Kalton, G. & Kish, L. (1981). Two efficient random imputation procedures. Proceedings of the Section on Survey Research Methods, ASA (146-153).

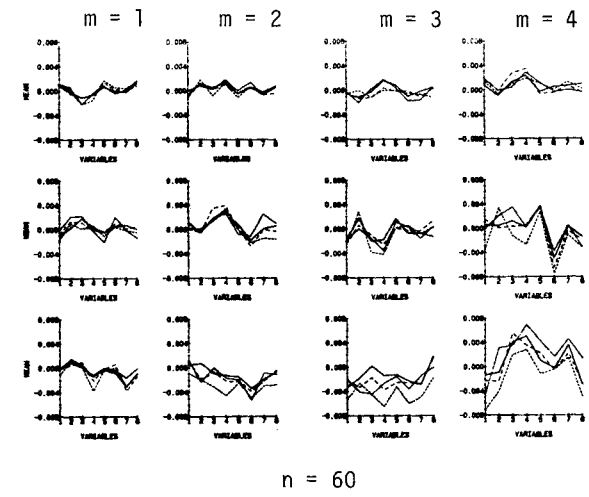
Oh, H. L. & Scheuren, F. J. (1980). Estimating the variance impact of missing CPS income data. Proceedings of the Section on Survey Research Methods, ASA (408-415).

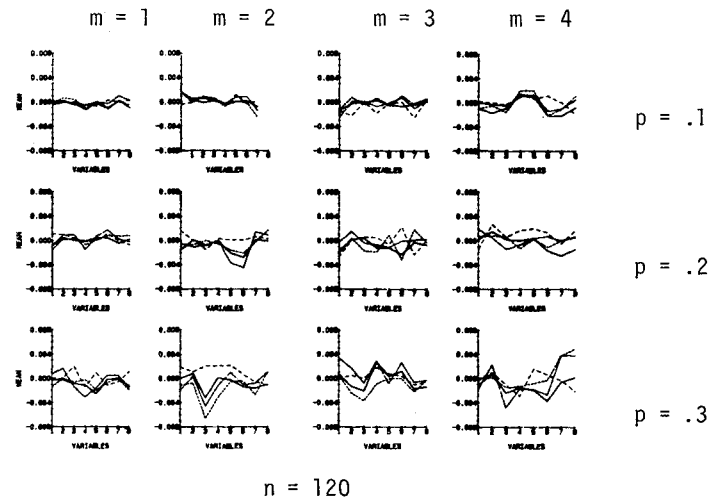
Proctor, C. H. (1978). More on imputing versus deleting when estimating scale scores. Proceedings of the Section on Survey Research Methods, ASA.

Schieber, S. J. (1978). A comparison of three alternative techniques for allocating unreported social security income on the survey of low-income aged and disabled. Proceedings of the Section on Survey Research Methods, ASA.

Timm, N. H. (1970) The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35:4 (417-437).





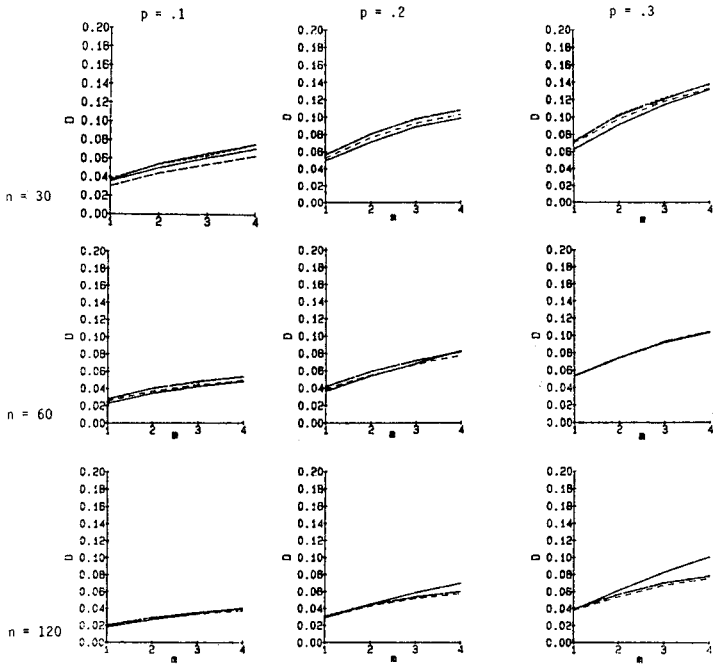


LEGEND

Hot-deck Random
Hot-deck Sequential
Hot-deck Distance
Cell Mean

DISCREPANCY IN MEANS OF COMPLETE AND IMPUTED SAMPLES AT VARIOUS LEVELS OF n, p, AND m

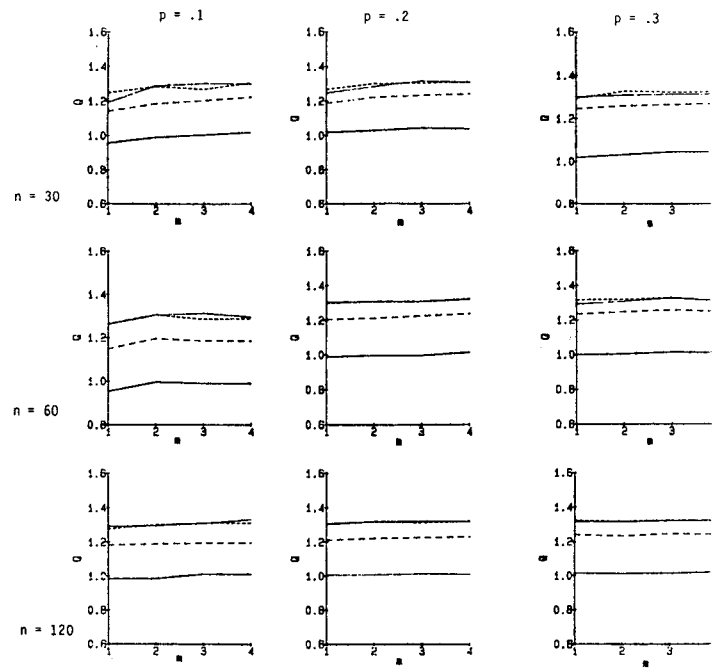
Figure 1



LEGEND: Hot-deck Random _____ Hot-deck Distance - - - - - -
Hot-deck Sequential ..... Cell Mean

RETENTION OF COMPLETE SAMPLE COVARIANCE STRUCTURE
IN IMPUTED SAMPLES AT VARIOUS LEVELS OF n, p, AND m

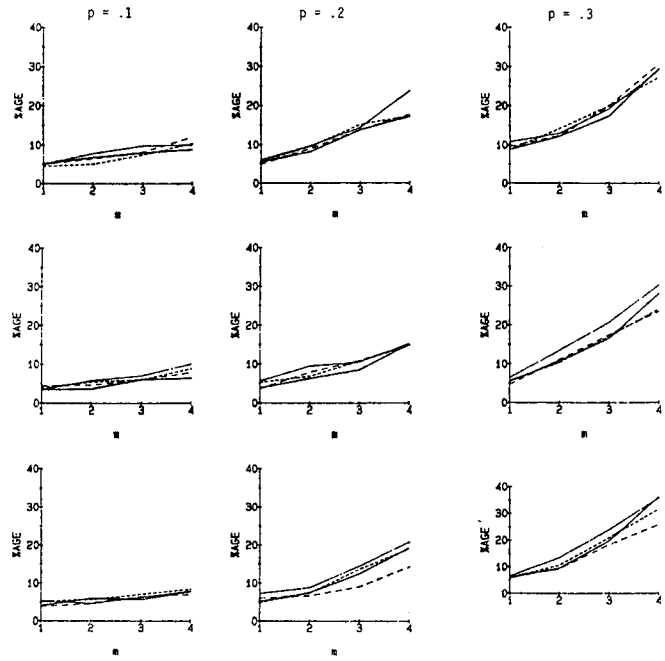
Figure 2



LEGEND: Hot-deck Random _____ Hot-deck Distance - - - - - -
Hot-deck Sequential ..... Cell Mean

QUALITY OF MISSING VALUE ESTIMATES
AT VARIOUS LEVELS OF n, p, AND m

Figure 3



LEGEND: Hot-deck Random _____ Hot-deck Distance - - - - - -
Hot-deck Sequential ..... Cell Mean

PERCENT OF IMPUTED SAMPLES THAT RETAINED POPULATION
COVARIANCE STRUCTURE AT VARIOUS LEVELS OF n, p, AND m

Figure 4