BIBLIOGRAPHY OF MULTIVARIATE PROCEDURES FOR UNBALANCED DATA

B. R. Murty, and W. T. Federer
Expert, Int'l. Atomic Energy Agency, Biometrics Unit
Maracaibo, Cornell University
Venezuela Ithaca, New York

BU-865-MA* December, 1984
(Rev. 5/85 & 12/86)

*In the Technical Report Series of the Biometrics Unit, Cornell
University, Ithaca, New York 14853.
BIBLIOGRAPHY OF MULTIVARIATE PROCEDURES FOR UNBALANCED DATA

B. R. Murty, Expert, W. T. Federer
Int'l. Atomic Energy Agency Biometrics Unit
Maracaibo, Cornell University
Venezuela Ithaca, New York

Key words:

ABSTRACT

Unbalanced data in multivariate situations cause considerable difficulties in performing statistical analyses. Many attempts for handling the missing data situation have been made, with varying degrees of success. In order to ascertain the current status of the situation, a bibliographic search is required. This is done in the present paper and is mostly confined to the last ten-year period.

INTRODUCTION

Unbalanced data in the real world is a frequently occurring phenomenon in experimentation. This is true for multivariate just as for univariate situations. Considerable effort has been expended to obtain statistical analyses for unbalanced univariate situations and accounts for a goodly portion of the area known as linear model theory. The situation for multivariate analysis is much different. Very little (relative to univariate) has been accomplished in providing statistical procedures for the general unbalanced multivariate situation.
Unbalanced data situations in a multivariate investigation fall into three categories, i.e.,

(i) data missing at random,
(ii) data missing by design and/or censored, and
(iii) data are impossible to obtain.

The procedures developed will be different for each situation. To ascertain what has been done, a search of the literature over the last ten years was made. The references were put into the following categories with number of citations listed in parentheses:

1. Effect of missing values on the multivariate normal assumptions (32).
2. Estimation of missing values (predictive) missing at random (33).
3. Estimation of missing values under censoring (17).
4. Missing values when data are impossible to obtain or are unobservable (1).
5. Missing values for multivariate discrete data (7).
6. Distribution-free (nonparametric) procedures for missing observations (8).
7. Testing hypotheses in the presence of missing data (26).

The bibliographic retrieval system known as BRS/After Dark Databases was used in the computer search. The key words used were multivariate analysis (missing observations or unbalanced or hierarchical). The Mathematical Reviews database which contains most of the major statistical journals was searched. Not all of the references contained a complete citation. The Current Index to Statistics was used to complete the reference where possible and was used to obtain additional references. The search was confined to be mainly within the last ten years.

The citations found are listed under each of the categories. It should be noted that very few or no references were found for a number of the categories. This attempt of categorizing the references will be useful to highlight the areas where the
available information or methodology is inadequate. This search revealed that in the last two years there was some activity on normality assumption (N), considerable activity on estimation of missing values under censoring (PC), and considerable activity on testing hypotheses in the presence of missing observations (T).

**Effect of Missing Values on Multivariate Normal Assumptions (N)**

All references listed below except for N were obtained from the literature search on the computer. There are 32 references on the effect of missing observations on the normality assumption. Among the 32 publications, 21 are in journals of statistics, while the others are distributed in miscellaneous publications as research reports and in mathematics journals. This indicates that the statisticians are aware of the basic problem of missing values on the validity of the main assumption of multivariate normality on which all the present tests of hypotheses and inferences are based. While the invalidity of normality assumption strikes at the very root of the present procedures in the interpretation of the data in applied situations, adequate effort for suitable modification of the procedures of analysis including transformation of the data to an approximate normality is not evident in the publications. The estimation procedures for the prediction of missing values are also based on multivariate normality assumption. Therefore, the solution for data not conforming to multivariate normality due to missing values should be the first priority area of further work.


N23 Radhakrishnan, R. (1982). Inadmissibility of the maximum likelihood estimator for a multivariate normal distribution when some observations are missing. *Comm. in Statist., A*, 11, 941-955. (See also under PR and T).


Estimation of Missing Values (Prediction) Missing at Random (PR)

Various methods of estimating missing values from a vector of observations have been put forward. They range all the way from omitting the vector to univariate methods to regression and covariance techniques. The methods all make use of the assumption that the observations are missing at random and that no selectivity is involved in their omission. Seven of the references came from Current Index to Statistics and the other 26 from the computer search. The total of 33 publications on prediction when the values are missing at random (PR) appears reasonable but far from the reality as most of the missing values are of a nonrandom nature in the real world. Twenty-one of the above 33 papers are in theoretical journals, 12 are in applied journals like Psychometrika or Technometrika; none are related to biological problems found in agriculture or animal production where missing data are very common. Thus, attention is paid more to a hypothetical situation where the missing observations are of a random nature.


PR9 Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. Amer. Statist. 36, 378-381. (See also under G.)


PR16 Greenlees, J. S. *et al.* (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *JASA* 77, 251-261.


**Estimation of Missing Values under Censoring (PC)**

In many situations the data will be censored. Different procedures will be needed from those when data are missing at random. Seventeen references were found, five by the computer search. Perhaps different key words such as multivariate analysis and censoring should have been used in the computer search. The problem of missing data under censoring in multivariate analysis is discussed in only 17 papers, 15 in theoretical publications and two in applied journals. The procedure followed in linear normal models when direct finding of missing values and computation of sufficient statistics is done
using "filled-in values" is not particularly appropriate in multivariate normal situations (see Dempster, Laird and Rubin, 1977). It is evident from the other papers that data from repeated sampling are often reported in censored form due to various reasons. Moreover, such censoring need not remain constant across sampling units. The complexity of the problem of censoring in multivariate analysis is brought out by Nelder and by Turnbull in their comments on the very interesting paper by Dempster, Laird and Rubin (1977). The problem is best summarized by Professor Turnbull: "In many problems it is hard to justify the assumption that the censoring mechanism is independent of the data observed or unobservable."


Missing Values when Data Are Impossible to Obtain or Are Unobservable (PU)

This is a very common situation met in biology as for some variables of subcellular characteristics in ultra-structural studies using electron microscopy. There is only one paper under this category and even that paper mentions only the problem. This aspect should receive immediate attention in future work.


Missing Values for Multivariate Discrete Data (D)

Seven references were found for this classification of missing observations from a multivariate vector of observations. However, some publications of a general nature refer to the need for studies of this problem for discrete data as is pointed out in discussions on the following pages.


D6: Mirkin, B. G. (1976). Analysis of quantitative tests (Mathematical models and methods) *Statistika, Moscow*, 1-166. (See also under G.)

All the seven papers listed above are in theoretical journals and indicate only the need to develop procedures for discrete data with missing values, as most of the present multivariate techniques are developed for continuous variables. Gnanadesikan and Kettenring (1984) reiterated that it is not difficult to deal with this problem. If such techniques are available, multiresponse studies and multi-way contingency data can be analyzed for their structure even with missing data.

**Distribution-Free (Nonparametric) Procedures for Missing Observations (NP)**

Only eight references were found which considered nonparametric procedures for multivariate situations with data missing at random. Among the publications on this aspect, all in theoretical journals, the one by Klotz (1980) is of considerable interest as it provides a modified Cochran-Friedman test procedure for testing equality of treatment means and also to construct a linear combination of treatments similar to those of single degree of freedom contrasts in the univariate case. The paper by Hanley and Parnes (1983) is equally useful in providing a procedure to construct a multivariate empirical survival function (MESF) from data even with heterogeneous censoring. Probably further work on this aspect will be forthcoming using iterative procedures like the EM algorithm.


Testing Hypotheses in the Presence of Missing Observations (T)

Eight of the 26 references on testing hypotheses with randomly missing data were found with the computer search. The others came from the Current Index to Statistics. All papers on this aspect are in theoretical journals, except for two or three. This is in contrast to the large number of papers concerning only prediction of missing values (see categories 2, 3, and 4). However, all the papers on testing hypotheses assume that the multivariate normality assumption is satisfied.

T1 Cohen, A. (1977). A result on hypothesis testing for a multivariate normal distribution when some observations are missing. J. Multivariate Anal. 7(3), 454-460.


**General (G)**

Seventy-four references were found which were believed to be useful or related to the missing data problem in multivariate analyses. Most of these were turned up in the computer search.
References G67 to G74 as well as 28 others were not found in the Current Index to Statistics. These references are somewhat incomplete. Among the 74 publications in this category, significant contributions are from the nine papers in the Series in Statistics and Probability by North-Holland Publishing Co., Amsterdam. Twenty-four more are from other theoretical statistics journals, 13 more in applied statistical journals, and the rest in diverse publications. There does not appear to be any study in depth by the theoretical statisticians for the solution of the general problems of missing values. Even the activity of the North-Holland group of publications is only a beginning.


G18 Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *Amer. Statist.* 38, 378-381. (See also under PR.)


G21 Engelman, L. (1982). An efficient algorithm for computing covariance matrices from data with missing values. *Comm. in Statist.*, B, 11, 113-121. (See also under PR.)


CONCLUSIONS

From the above classification of the available recent publications in relation to multivariate analysis with missing data, it is clear that large gaps exist in our knowledge in all the areas where practical needs exist, particularly about the effecting of missing values on the multivariate normality assumption and in the estimation procedures when the missing values are unobservable or are due to censoring. In the case of testing of hypotheses, the few available papers concentrate on values missing at random. There is little effort in the area of multivariate discrete data with missing values, although it is considered easy to extend the procedures of continuous variables to discrete variables, with some modification. The development of distribution-free procedures in this field is still in the initial stages. A concentrated effort in this area will help overcome many of the problems of multivariate normality assumption, nonrandom nature of missing values, mixtures of distributions, and simultaneous multivariate analysis of data with continuous and discrete variables. As emphasized by Gnanadesikan and Kettenring (1984) so far "there is little evidence of matching the method to the real needs of the problem" and there is "the tendency of the user's willingness to settle for the routine output of the method." With the recognition of the need, an accelerated effort to remove the gaps is worthwhile for the extended and proper use of multivariate analysis with missing values in applied research.