BIBLIOGRAPHY OF MULTIVARIATE PROCEDURES FOR UNBALANCED DATA

B. R. Murty,
Expert, Int'l. Atomic Energy Agency,
Maracaibo,
Venezuela

and

W. T. Federer
Biometrics Unit
Cornell University
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,
Int'l. Atomic Energy Agency
Maracaibo,
Venezuela

W. T. Federer
Biometrics Unit
Cornell University
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
8. General (74).

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.

N1 Boyles, R. A. and Samaniego, F. J. (1984). Modeling and inference for multivariate binary data with positive dependence. *J. Amer. Statist. Assoc.* **79**, 188-193.

N2 Bock, H. H. (1979). Mathematisch statistische Modelle in der Cluster analyse. *Physica*, Wurzburg, Germany.

N3 Brailovsky, V. (1980). On the influence of missing data in a sample set on the quality of a statistical decision rule. *New York Acad. Sci.,* New York.

N4 Brailovsky, V. (1981). Remark concerning a paradoxical situation in behaviour of error rate in discriminant analysis. *Statist. Anal. Donnees* **6**(1), 28-38.

N5 Brown, C. H. (1983). Asymptotic comparison of missing data procedures for estimating factor loadings. *Psychometrika* **48**, 269-291.

N6 Conway, D. and Theil, H. (1980). The maximum entropy moment matrix with missing values. *Econom. Lett.* **5**(4), 319-322.

N7 Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8**(3), 522-539.

N8 Dempster, A. P., Laird, N. M., and Rubin, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. *Multivariate Analysis* **5**, 35-57.

N9 Gnanadesikan, R., Kettenring, J. R., and Landwehr, J. M. (1977). Interpreting and assessing the results of cluster analyses. *Bull. Inst. Internat. Statist.* **47**(2), 451-493.

N10 Gupta, A. K. and Rohatgi, V. K. (1982). Estimation of covariance from unbalanced data. *Sankhya, Ser. B.* **44**(2), 143-153. (See also under T.)

N11 Hathaway, R. J. (1986). Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters* **4**, 53-56.

N12 Ito, K. (1962). A comparison of the powers of two multivariate analysis of variance tests. *Biometrika*, **50**, 455-462. (An interesting paper.)

N13 Ito, K. (1966). On the heteroscedasticity in the linear normal regression model. In *Res. Papers Statist. Festchr., Neyman* (F. David, ed.), 147-155. (A useful paper on role of sample size on multivariate normality assumption.)

N14 James, A. T. and Venables, W. (1980). Interval estimates for a bivariate principal axis. *Multivariate Analysis* **5**, 399-411.

N15 Kabe, D. G. (1981). Some results for the univariate normal random linear regression model prediction theory. U. S. Army Res. Office, Research Triangle Park, NC.

N16 Kawasaki, S. and Zimmermann, K. F. (1981). Measuring relationships in the log-linear probability model by some compact measures of association. *Statist. Hefte (N.F.)* **22**(2), 82-109.

N17 Machado, S. G. (1983). Two statistics for testing multivariate normality. *Biometrika* **70**(3), 713-718.

N18 Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. *Comm. Statist. A — Theory & Methods* **7**(13), 1233-1241.

N19 Mardia, K. V. (1984). Mardia's test of multinormality. Unpublished manuscript.

N20 Murray, G. D. (1979). The estimation of multivariate normal density functions using incomplete data. *Biometrika* **66**(2), 375-380.

N21 Niemi, H. and Weron, A. (1981). Dilation theorems for positive definite operator kernels having majorants. *J. Funct. Anal.* **40**(1), 54-65.

N22 Payne, R. W. (1981). Selection criteria for the construction of efficient diagnostic keys. *J. Statist. Plan. Inference* **5**(1), 27-36.

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).

N24 Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *Siam Review* **26**, 195-202.

N25 Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.* **79**, 406-414.

N26 Springer, M. D. (1979). *The algebra of random variables.* John Wiley & Sons, New York, Chichester, Brisbane. xix + 470 pp.

N27 Srivastava, J. N. and McDonald, L. L. (1974). Analysis of growth curves under the hierarchical models. *Sankhyā, Ser. A,* **36**(3), 251-260.

N28 Szatrowski, T. H. (1985). Asymptotic distributions in the testing and estimation of the missing-data multivariate normal linear patterned mean and correlation matrix. *Linear Algebra and Its Applications* **67**, 215-231.

N29 Titterington, D. M. (1977). Analysis of incomplete multivariate binary data by the kernel method. *Biometrika* **64**(3), 455-460. (See also under D and NP.)

N30 Walker, J. J. (1979). An asymptotic expansion for unbalanced quadratic forms in normal variables. *J. Amer. Statist. Assoc.* **74**, 389-392.

N31 Yakowitz, S. J. and Szidarovszky, F. (1985). A comparison of kriging with nonparametric regression methods. *J. of Multivariate Anal.* **16**, 21-53.

N32 *Proceedings of the Second International Conference on Mathematical Modelling, Volumes I & II.* University of Missouri at Rolla, Rolla, MO.

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.

PR1 Agresti, A. (1981). A hierarchical system of interaction measures for multidimensional contingency tables. *J. Roy. Statist. Soc., Ser. B,* **43**(3), 293-301.

PR2 Anderson, O. (Editor). (1982). Time series analysis: theory and practice. 1. *Proceedings, International Conference, Valencia, Spain,* North-Holland Publishing Co., Amsterdam, New York. ix + 756 pp. (See also under PC.)

PR3 Bailey, T. A., Jr. and Dubes, R. Cluster validity profiles. *Pattern Recognition* **15**(2), 61-83.

PR4 Basu, D. and Pereira, C. A. de B. (1982). On the Bayesian analysis of categorical data: the problem of nonresponse. *J. Statist. Plan. Inference* **6**(4), 345-362.

PR5 Batagelj, V. (1979). Note on: "Ultrametric hierarchical clustering algorithms." *Psychometrika* **44**(3), 343-346.

PR6 Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *J. Roy. Statist. Soc. Ser. B* **37**, 129-145. (See also under PC.)

PR7 Braverman, E. L., Levina, A. I., and Levin, M. I. (1980). Hierarchical aggregation of multiple classifications. *Automat. Remote Control* **41**(6, part 2), 848-853.

PR8 Dahiya, R. C. and Korwar, R. M. (1980). Maximum likelihood estimates for a bivariate normal distribution with missing data. *Ann. Statist.* **8**(3), 687-692.

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.)

PR10 Engelman, L. (1982). An efficient algorithm for computing covariance matrices from data with missing values. *Comm. Statist., Ser. B,* **11**, 113-121. (See also under G.)

PR11 Enke, H. (1978). On the analysis of special incomplete three-dimensional contingency tables. *Biometrical J.* **20**(3), 229-242.

PR12 Ferligoj, A. and Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika* **47**(4), 413-426.

PR13 Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika* **41**(3), 409-415.

PR14 Giguere, M. A. and Styan, G. P. H. (1975). Comparisons between maximum likelihood and naive estimators in a multivariate normal population with data missing on one variate. *Bull. Inst. Int'l. Statist.* **40**(3), 303-308.

PR15 Giguere, M. A. and Styan, G. P. H. (1978). *Multivariate Normal Estimation with Missing Data on Several Variates.* Academia, Prague.

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.

PR17 Hosking, J. D. (1981). Missing data in multivariate linear models: A comparison of several estimation techniques. SAS *SUGI* **6**, 46-51.

PR18 Koornwinder, T. (1978). Positivity proofs for linearization and connection coefficients of orthogonal polynomials satisfying an addition formula. *J. London Math. Soc.* **18**(1), 101-114.

PR19 Korn, E. L. (1981). Hierarchical log-linear models not preserved by classification error. *JASA* **76**, 110-113.

PR20 Linares, G. and Mederos, M. V. (1982). The reduced model and the method of Tocher in parameter estimation with missing observations in regression models. *Cienc. Mat. (Havana)* **3**(1), 107-127.

PR21 Little, R. J. A. (1979). Maximum likelihood inference for multiple regression with missing values: a simulation study. *J. Roy. Statist. Soc., Ser. B,* **41**(1), 76-87.

PR22 Lukasova, A. (1979). Hierarchical agglomerative clustering procedure. *Pattern Recognition* **11**(5-6), 365-381.

PR23 Radhakrishnan, R. (1982). Inadmissibility of the maximum likelihood estimator for a multivariate normal distribution when some observations are missing. *Comm. Statist., Ser. A - Theory & Methods* **11**(8), 941-955. (See also under N and T.)

PR24 Sarkar, S. K. (1979a). A test for mean with additional observations. *Calcutta Statist. Assoc. Bull.* **28**(109-112), 47-56.

PR25 Sarkar, S. K. (1979b). On optimum properties of a likelihood ratio test with additional information. *Sankhyā, Ser. A,* **41**(3-4), 207-218.

PR26 Sarkar, S. K., Sinha, B. K., and Krishnaiah, P. R. (1983). Some tests with unbalanced data from a bivariate normal population. *Ann Inst. Statist. Math.* **35**(1), 63-75.

PR27 Schader, M. Hierarchical analysis: classification with ordinal object dissimilarities. *Metrika* **27**(2), 127-132.

PR28 Smith, P. L. (1981). The use of analysis of covariance to analyse data from designed experiments with missing or mixed-up values. *Appl. Statist.* **30**, 1-8.

PR29 Stewart, W. E. and Soerensen, J. P. (1981). Bayesian estimation of common parameters from multiresponse data with missing observations. *Technometrics* **23**(2), 131-141.

PR30 Stewart, W. E. and Soerensen, J. P. (1982). "Correction to Bayesian estimation of common parameters from multiresponse data with missing observations." *Technometrics* **24**, 91.

PR31 Stoffer, D. S. (1986). Estimation and identification of space-time Armax models in the presence of missing data. *J. Amer. Statist. Assoc.* **81**, 762-772.

PR32 Timm, N. H. (1980). The analysis of nonorthogonal MANOVA designs employing a restricted full rank multivariate linear model. In *Multivariate Statistical Analysis,* (R. P. Gupta, ed.). North-Holland, Amsterdam.

PR33 Titterington, D. M., Murray, G. D., *et al.* (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *J. Roy. Statist. Soc., Ser. A,* 144(2), 145-175 (with discussion).

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."

PC1 Anderson, O. D. (1982). Time series analysis: theory and practice. 2. *Proceedings, Int'l. Conference, Dublin, Ireland.* North-Holland Publishing Co., Amsterdam, New York. viii + 756 pp. (See also under PR.)

PC2 Andrade, D. F. and Helms, R. W. (1984). Maximum-likelihood estimates in the multivariate normal with patterned mean and covariance via the EM algorithm. *Comm. in Statist., Theory and Methods A* **13**, 2239-2251.

PC3 Basilevsky, A., Sabourin, D., Huns, D., and Anderson, A. (1985). Missing data estimators in the general linear model: An evaluation of simulated data as an experimental design. *Comm. in Statist., B* **14**, 371-394.

PC4 Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *J. Roy. Statist. Soc. B* 37, 129-145. (See also under PR.)

PC5 Bruynooghe, M. (1980). An accelerated clustering algorithm based on the convex hull concept. *COMPSTAT* **4**, 412-418.

PC6 Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc., Ser. B,* **39**, 1-38. (With discussion) (See also under PU.)

PC7 Desarbo, W. S., Green, P. E., and Carroll, J. D. (1986). An alternating least-squares procedure for estimating missing preference data in product-concept testing. *Decision Sciences* **17**, 163-185.

PC8 Hosking, J. D. (1984). A comparison of several procedures for estimation in incomplete multivariate linear models. *Proceedings, Joint Statistical Meetings of ASA and Biometric Society*, August 13-16. 245 pp.

PC9 Iwasaki, K., Yoshioka, N., and Matoba, Y. (1978). Evaluation of hierarchical clustering techniques by the criterion functions for partition. *Systems Comput. Controls* **8**(1), 25-32.

PC10 Kariya, T., Krishnaiah, P. R., and Rao, C. R. (1983). Inference on parameters of multivariate normal population when data are missing. In *Developments in Statistics, Vol. 4* (P. R. Krishnaiah, ed.) Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers), New York, London. 137-184.

PC11 Lutkepohl, H. (1986). Forecasting vector arma processes with systematically missing observations. *J. Business & Economic Statist.* **4**, 375-390.

PC12 Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *JASA* **79**(2), 406-414.

PC13 Rubin, D. B. and Szatrowski, T. H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika* **69**(3), 657-660.

PC14 Shih, W. J. and Weisberg, S. (1986). Assessing influence in multiple regression with incomplete data. *Technometrics* **28**, 231-239.

PC15 Simon, G. A. and Simonoff, J. S. (1986). Diagnostic plots for missing data in least-squares regression. *J. Amer. Statist. Assoc.* **81**, 501-509.

PC16 Szatrowski, T. H. (1983). Missing data in the one-population multivariate normal patterned mean and covariance-matrix testing and estimation problem. *Ann. Statist.* **11**, 947-958.

PC17 Titterington, D. M. (1983). Kernel-based density-estimation using censored, truncated or grouped data. *Comm. in Statist., Theory and Methods A* **12**, 2151-2167.

PC18 Zeger, S. L. and Brookmeyer, R. (1986). Regression-analysis with censored autocorrelated data. *J. Amer. Statist. Assoc.* **81**, 722-729.

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.

PU1 Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc., Ser. B,* **39**, 1-38 (with discussion). (See under PC.)

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.

D1: Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis Theory and Practice.* MIT Press, Cambridge, Mass.

D2: Burbla, J. and Rao, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces. A unified approach. *J. Multivariate Analysis* **12**, 575-596.

D3 Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc., B,* **39**, 1-38. (With discussion) (See under PU, PC.)

D4: Gnanadesikan, R. and J. R. Kettenring (1984). A pragmatic view of multivariate methods in applications. In *Statistics: An Appraisal*, (H.A. David and H.T. David, eds.), Iowa State University Press, Ames, Iowa, 309-337.

D5: Goodman, L. A. (1978). *Analyzing Qualitative/Categorical Data.* Abbott Books, Cambridge, Mass., viii + 471 pp. (See also under G.)

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

D7: Titterington, D.M. (1977). Analysis of incomplete multivariate binary data by the kernel method. *Biometrika,* 64(3), 455-460. (See also under N and NP.)

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, multi-response 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.

NP1 Ambrosi, K. (1979). A distribution-free procedure in discriminant analysis with arbitrarily structure attributes. *Hain,* Meisenheim.

NP2 Basu, A. P., Ghosh, J. K., and Sen, P. K. (1983). A unified way of deriving LMP rank-tests from censored-data. *J. Roy. Statist. Soc. B* **45**, 384-390.

NP3 Hanley, J. A. and Parnes, M. N. (1983). Nonparametric estimation of a multivariate distribution in the presence of censoring. *Biometrics,* **39**, 129-139.

NP4 Klotz, J. (1980). A modified Cochran-Friedman test with missing observations and ordered categorical data. *Biometrics* **36**(4), 665-670.

NP5 Laird, N. M. (1976). Nonparametric maximum likelihood estimation of a distribution function with mixtures of distributions. Technical Report S-47, NS-338, Dept. of Statistics, Harvard University, Cambridge, Mass.

NP6 Papaioannou, T. and Speevak, T. (1977). Rank correlation inequalities with missing data. *Comm. Statist. - Theory & Methods* A **6**(1), 67-72.

NP7 Titterington, D. M. (1977). Analysis of incomplete multivariate binary data by the kernel method. *Biometrika* **64**(3), 455-460. (See also under D and N.)

NP8 Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *J. Amer. Statist. Assoc.* **79**, 653-661.

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.

T2 Das, K. (1978). BAN estimators of variance components for the unbalanced one-way classification. *Calcutta Statist. Assoc. Bull.* **27**(105-108), 97-118.

T3 Eaton, M. and Kariya, T. (1983). Multivariate tests with incomplete data. *Ann. Statist.* **11**, 654-665.

T4 Giguere, M. A. and Styan, G. P. H. (1975). Comparisons between maximum likelihood and naive estimators in a multivariate normal population with data missing on one variate. *Proceedings, Bull. Int'l. Statist. Inst.* **40**, 303-308 (with discussion).

T5 Green, P. J. (1984). Iteratively reweighted least-squares for maximum-likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. B* **46**, 149-192 (with discussion).

T6 Gupta, A. K. and Rohatgi, V. K. (1982). Estimation of covariance from unbalanced data. *Sankhyā, Ser. B,* **44**(2), 143-153. (See also under N.)

T7 Kaplan, J. (1982). A theorem relating MINQUE and unweighted means estimators of variance components in the one-way design. *Comm. Statist., Ser. A — Theory & Methods* **11**(4), 423-428.

T8 Kariya, T. (1981). Robustness of multivariate tests. *Ann. Statist.* **9**(6), 1267-1275.

T9 Kariya, T. (1985). A new concept of second order efficiency and its application to a missing data problem. *Statist. Theory & Decision Anal.* (Editor, K. Matusita), pp. 331-353.

T10 Kariya, T., Krishnaiah, P. R., and Rao, C. R. (1983). Inference on parameters of multivariate normal populations when some data is missing. In *Developments in Statistics, Vol.4,* (P.R. Krishnaiah, ed.). Academic Press, Inc. (Harcourt Brace Jovanovich, Publishers), New York, London. pp. 137- 184.

T11 Koopman, R. F. (1978). On Bayesian estimation in unrestricted factor analysis. *Psychometrika* **43**(1), 109-110.

T12 LaMotte, L. R. (1980). Some results on biased linear estimation applied to variance component estimation. *Mathematical Statistics and Probability: Proc. Sixth Int'l. Conf., Wisla, Poland,* Springer, Berlin.

T13 Little, R. J. A. (1976). Inference about means from incomplete multivariate data. *Biometrika,* **63**, 593-604.

T14 Little, R. J. A. (1979). Maximum likelihood inference for multiple regression with missing values: A simulation study. *J. Royal Statist. Soc., Ser. B,* **41**(1), 76-87. (See also under PR.)

T15 Little, R. J. A. and Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *Amer. Stat.* **37**, 218-220.

T16 Marden, J. I. (1981). Invariant tests on covariance matrices. *Ann. Statist.* **9**(6), 1258-1266.

T17 Nordheim, E. V. (1984). Inference from non-randomly missing categorical data. An example from a genetic study on Turner's syndrome. *J. Amer. Statist. Assoc.* **79**, 772-780.

T18 Orchard, T. and Woodbury, M. A. (1970). A missing information principle: Theory and applications. *Proc. 6th Berkeley Symp. on Math. Statist. & Probability* I, 697-713.

T19 Radhakrishnan, R. (1982). Inadmissibility of the maximum likelihood estimator for a multivariate normal distribution when some observations are missing. *Comm. Stat., Ser. A — Theory & Methods* **11**, 941-955. (See also under N and PR.)

T20 Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581-592.

T21 Smith, W. B. and Riggs, M. W. (1984). Likelihood ratio testing on partial multinormal data. *Statistics & Probability Letters* **2**, 337-343.

T22 Sneath, P. H. A. (1986). Significance tests for multivariate normality of clusters from branching patterns in dendrograms. *Mathematical Geology* **18**, 3-32.

T23 Srivastava, M. S. (1985). Multivariate data with missing observations. *Comm. in Statist. Theory and Methods A* **14**, 775-792.

T24 Szatrowski, T. H. (1985). Missing data in the k-population multivariate normal patterned mean and covariance matrix testing and estimation problem. *Comm. in Statist. B* **14**, 357-370.

T25 Szatrowski, T. H. (1985). Asymptotic distributions in the testing and estimation of the missing-data multivariate normal linear patterned mean and correlation. *Linear Algebra Ap.* **67**, 215-231.

T26 Wesolowska-Janczarek, M. T. (1984). Estimation of covariance matrices in unbalanced random and mixed multivariate models. *Biometrical J.* **26**, 665-674.

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.

G1 Albert, J. H. and Gupta, A. K. (1982). Mixtures of Dirichlet distributions and estimation in contingency tables. *Ann. Statist.* **10**(4), 1261-1268.

G2 Anderson, E. B. (1982). Latent structure analysis: a survey. *Scand. J. Statist.* **9**(1), 1-12.

G3 Anderson, T. W. (1984). *An Introduction to Multivariate Analysis*, 2nd ed., John Wiley & Sons, New York, pp 1-675.

G4 Atwood, C. L. (1986). The binomial failure rate common cause model. *Technometrics* **28**, 139-148.

G5 Baker, G. A., Jr. and Graves, M. P. (1981). *Pade approximants. Part II.* Addison-Wesley Publishing Co., Reading, Mass. vxiii + 215 pp.

G6 Basford, K. E. and McLachlan, G. J. (1985). Likelihood estimation with normal mixture-models. *Appl. Statist.* **34**, 282-289.

G7 Campbell, N. A. (1984). Missing value – canonical variate analysis – a general model formulation. *Australian J. Statist.* **26**, 86-96.

G8 Campbell, N. A. and Tomenson, J. A. (1983). Canonical variate analysis for several sets of data. *Biometrics* **39**(2), 425-435.

G9 Chandon, J. L. and Pinson, S. (1981). *Typological analysis.* Masson, Paris. x + 254 pp.

G10 Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Statist.* **32**, 267-275.

G11 Chen, T. T., Hochberg, Y., and Tenenbein, A. (1984). Analysis of multivariate categorical data with misclassification errors by triple sampling schemes. *J. Statist. Planning and Inference* **9**, 177-184.

G12 Clayton, D. and Cuzick, J. (1985). The EM algorithm for Cox regression model using GLIM. *Appl. Statist.* **34**, 148-156.

G13 Cox, C. (1984). Generalized linear models - the missing link. *Appl. Statist.* **33**, 18-24.

G14 Cox, M. A. A. and Plackett, R. L. (1980). Small samples in contingency tables. *Biometrika* **67**(1), 1-13.

G15 Crepeau, H., Koziol, J., Reid, N., and Yuh, Y. S. (1985). Analysis of incomplete multivariate data from repeated measurement experiments. *Biometrics* **41**, 505-514.

G16 DeGroot, M. H., Bernardo, J. M., Lindley, D. V., and Smith, A. F. M. (editors). (1979). *Bayesian Statistics. Proceedings First Int'l. Meeting, Valencia, Spain, May 28-June 2.* University Press, Valencia. 647 pp.

G17 Delattre, M. and Hansen, P. (1979). *Cluster analysis and graph coloring.* North-Holland, Amsterdam.

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.)

G19 Drygas, H. (1976). Gauss-Markov estimation for multivariate linear models with missing observations. *Ann. Statist.* **4**(4), 779-787.

G20 Dufour, J. M. and Dagenais, M. G. (1985). Durbin-Watson tests for serial correlation in regressions with missing observations. *J. Econometrics* **27**, 371-381.

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.)

G22 Ferrandiz, J. R. (1982). A Bayesian solution to the Stein's paradox. *Trabajos Estadist. Investigacion Oper.* **33**(2), 31-46.

G23 Fichet, B. (1981). On approximations of dissimilarity indices via Euclidean and hierarchical representations. *Statist. Anal. Donnees* **6**(2), 1-21.

G24 Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika* **44**(4), 409-420.

G25 Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations.* John Wiley & Sons, New York. 311 pp.

G26 Gondran, M. (1976). La structure algebrique des classifications hierarchiques. *Ann. I.N.S.E.E.,* No. 22-23, 181-190.

G27 Gondran, M. (1977). Eigenvalues and eigenvectors in hierarchical classifications. *Proceedings, Recent Developments in Statistics,* 1976 Meeting of European Statisticians, Sept. 6-11, Grenoble, France. North-Holland, Amsterdam. pp. 775-782.

G28 Gondran, M. (1980). Classification hierarchique et flots dans un graphe. *Bull. Direction Etudes Rech., Ser. C - Math. Inform.* 1-3, 43-48.

G29 Goodman, L. A. (1978). *Analyzing Qualitative/Categorical Data.* Abt Books, Cambridge, Mass. viii + 471 pp.

G30 Goodman, L. A. (1985). The analysis of cross-classified data having ordered and or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Statist.* **13**, 10-69.

G31 Grim, J. (1986). Multivariate statistical pattern-recognition with nonreduced dimensionality. *Kybernetika* **22**, 142-157.

G32 Harvey, A. C. and Peters, S. (1985). A note on the estimation of variances in state-space models using the maximum *a-posteriori* procedure. *Trans. Automatic Control* **30**, 1048-1050.

G33 Hubert, L. J. (1974). Some applications of graph theory to clustering. *Psychometrika* **39**, 283-309.

G34 Hubert, L. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *JASA* **69**, 698-704.

G35 Innes, G. and O'Neill, R. V. (1979). *Systems Analysis of Ecosystems.* International Co-operative Publishing House, Burtonsville, MD. xxvii + 402 pp.

G36 Jambu, M. and Lebeaux, M. O. (1978). *Automatic Classification for Data Analysis. II.* Dunod, Paris, 1978.

G37 James, A. T. (1982). *Analysis of Variance Determined by Symmetry and Combinatorial Properties of Zonal Polynomials.* North-Holland, Amsterdam.

G38 Johnson, R. A. and Wichern, D. W. (1982). *Applied Multivariate Statistical Analysis.* Prentice-Hall, Inc., Englewood Cliffs, NJ. xiii + 594 pp.

G39 Kabe, D. G. (1980). Direct solutions to the m-median and fractional transportation problems. *Indust. Math.* **30**(1), 1-27.

G40 Kochevar, P. (1984). An application of multivariate B-splines to computer aided geometric design. *Rocky Mountain J. Math.* **14**, 159-175.

G41 Kouvatsos, D. D. (1977). Decomposition criteria for complex systems with reference to computer design. *Hemisphere,* Washington, DC.

G42 Krishnaiah, P. R. (Ed.) (1980). *Analysis of Variance.* North-Holland Publishing Co., Amsterdam, New York. xvii + 1002 pp.

G43 Krivjakova, E. N. (1978). An omega 2-type criterion for testing simple and composite hypotheses in the multivariate case. *"Nauka" Sibirsk. Otdel.,* Novosibirsk.

G44 Kshirasagar, A. M. (1972). *Multivariate Analysis*, Marcel Dekker Inc., New York. pp. 1-533.

G45 Laird, N. (1985). Missing information principle. *Enc. Statist. Sci.* **5**, 548-552.

G46 Landis, J. R., Heyman, E. R., and Koch, G. G. (1978). Average partial association in three-way contingency tables: a review and discussion of alternative tests. *Int'l. Statist. Rev.* **46**(3), 237-254.

G47 Leclerc, B. (1977). An application of combinatorial theory to hierarchical classification. In *Development of Statistics* North-Holland, Amsterdam, pp. 783-786.

G48 Liski, E. P. (1985). Estimation from incomplete data in growth curves models. *Comm. in Statist., Simulation & Computation, B* **14**, 13-27.

G49 Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497-512.

G50 Martin, R. J. (1984). Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Comm. in Statist., Theory & Methods A* **13**, 1275-1288.

G51 Meyer, K. (1985). Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics* **41**, 153-165.

G52 Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. *Psychometrika* **44**(3), 343-346.

G53 Mirkin, B. G. (1976). Analysis of quantitative tests (mathematical models and methods). Izdat. Statistika, Moscow. 166 pp. (See also under D.)

G54 Mojena, R. and Wishart, D. (1980). Stopping rules for Ward's clustering method. *COMPSTAT.* **4**, 426-432.

G55 Mudholkar, G. S. and Subbaiah, P. (1980). *A Review of Step-Down Procedures for Multivariate Analysis of Variance.* North-Holland, Amsterdam.

G56 Ohsumi, N. and Nakamura, T. (1981). Some properties of monotone hierarchical dendrogram in numerical classifications. *Proceedings Inst. Statist. Math.* **28**(1), 117-133.

G57 Orchard, T. and Woodbury, M. A. (1970). A missing information principle: Theory and applications. *Proc. 6th Berkeley Symp. on Math. Statist. and Probabilities*, Vol. I, 697-713.

G58 Sibson, R. (1984). Present position and present development: Some personal views. Multivariate analysis. *J. Roy. Statist. Soc. A* **147**, 198-207 (with discussion).

G59 Smith, P. L. (1981). The use of analysis of covariance to analyze data from designed experiments with missing or mixed-up values. *Appl. Statist.,* **30**, 1-8.

G60 Smithson, M. (1982). Applications of fuzzy set concepts to behavioral sciences. *Math. Social Sci.* **2**(3), 257-274.

G61 Srivastava, M. S. (1985). Multivariate data with missing observations. *Comm. in Statist., Theory & Methods A* **14**, 775-792.

G62 Tidmore, F. E. and Turner, D. W. (1983). On clustering with Chernoff-type faces. *Comm. Statist., Ser. A – Theory & Methods* **12**(4), 381-396.

G63 Warren, E., Stewart, J., and Sorenson, P. (1981). Bayesian estimation of common parameters from multi-response data with missing observations. *Technometrics* **23**, 131-141.

G64 Weiss, H. R. (1978). *Approximate and Exact Tests for the Analysis of Multidimensional Contingency Tables.* Physica Verlag, Wurzburg. 136 pp.

G65 Weiss, M. C. (1978). Decomposition hierarchique du Khi-deux associe a un tableau de contingence a plusieurs entrees. *Rev. Statist. Appl.* **26**(1), 23-33.

G66 Zacks, S. and Rodriguez, J. (1986). A note on the missing value principle and the EM-algorithm for estimation and prediction in sampling from finite populations with a multinormal superpopulation model. *Statist. & Prob. Letters* **4**, 35-37.

G67 *Classification and Clustering.* (1980). "Mir," Moscow. 389 pp.

G68 *Essays in Probability and Statistics.* (1976). Published by the Editorial Committee for Publication of Essays in Probability and Statistics; distributed by Shinko Tsush & Co., Ltd., Tokyo. xix + 716 pp.

G69 *Identification and System Parameter Estimation. Part 3.* North-Holland Publishing Co., Amsterdam, New York. i-x + pp. 1367-2178.

G70 *Pattern Recognition in Practice.* North-Holland Publishing Co., Amsterdam, New York. xii + 552 pp.

G71 *Proceedings of the Econometric Society, 1979 European Meeting.* North-Holland Publishing Co., Amsterdam, New York. xv + 444 pp.

G72 *Third Formator Symposium on Mathematical Methods for the Analysis of Large-Scale Systems.* (1979). Academia (Publishing House of the Czechoslovak Academy of Sciences), Prague. 382 pp.

G73 *Time Series.* North-Holland Publishing Co., Amsterdam, New York. viii + 446 pp.

G74 *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes* and of the *Eighth European Meeting of Statisticians, Vol. B.* (1978).Academia (Publishing House of the Czechoslovak Academy of Sciences), Prague. 582 pp.

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.