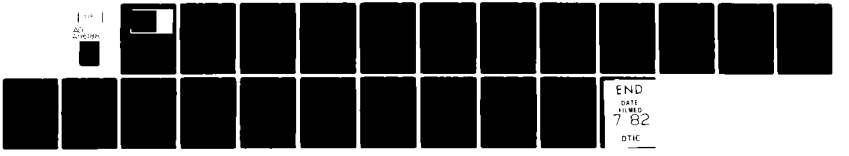
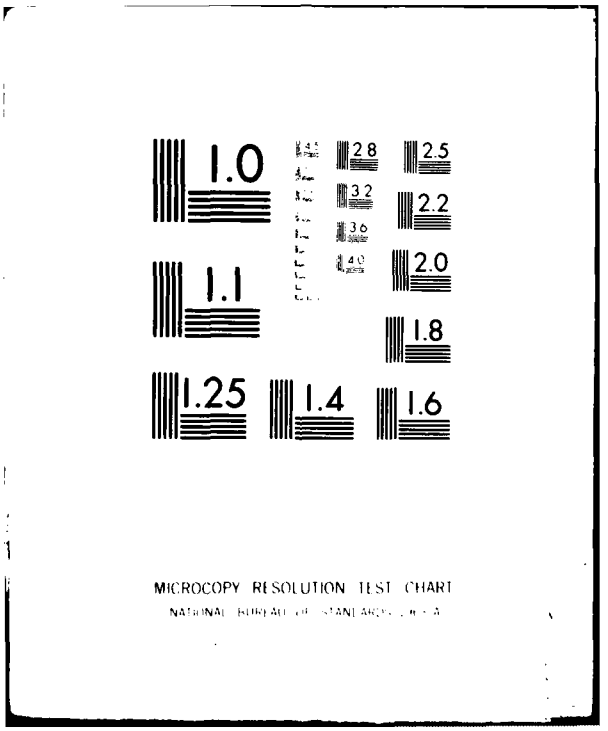
AD-A116 188 WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
INCOMPLETE DATA - ENCYCLOPEDIA ENTRY.(U)
APR 82 R J LITTLE, D B RUBIN DAAG29-80-C-0041
UNCLASSIFIED MRC-TSR-2372 NL



END
DATE FILMED
7 82
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART

MRC Technical Summary Report # 2372

INCOMPLETE DATA - ENCYCLOPEDIA ENTRY

Roderick J. A. Little
and
Donald B. Rubin

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

April 1982

(Received October 27, 1981)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

INCOMPLETE DATA - ENCYCLOPEDIA ENTRY

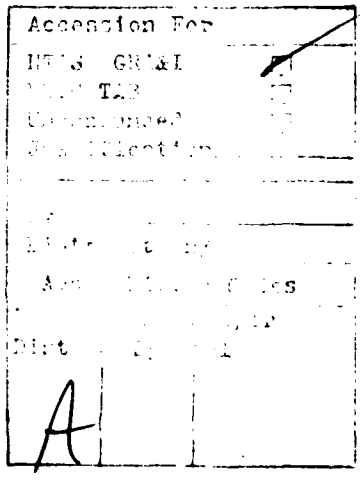
Roderick J. A. Little and Donald B. Rubin

Technical Summary Report #2372

April 1982

ABSTRACT

Incomplete data occur commonly in the application of statistics to real data, and there exists a substantial literature on the problem. This article provides an overview of issues for the statistically knowledgeable but not necessarily statistically sophisticated reader. It is intended to appear as an entry in The Encyclopedia of Statistical Sciences.



AMS (MOS) Subject Classifications: 6207, 62D05, 62H12

Key Words: Missing Data, EM Algorithm, Nonresponse, Censored Data, Truncated Data, Imputation, Factorizing Likelihoods

Work Unit Number 3 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

INCOMPLETE DATA - ENCYCLOPEDIA ENTRY

Roderick J. A. Little and Donald B. Rubin

1. Introduction

Incomplete data is an extremely general problem in statistics. Indeed, one might view inferential statistics in general as a collection of methods for extending inferences from a sample to a population where the non-sampled values are regarded as missing data.

Although some statistical methods for complete data, such as factor analysis, finite mixture models, and mixed model analysis of variance can be usefully viewed as incomplete data methods (Dempster, Laird and Rubin, 1977), we restrict this review to more standard incomplete data problems. For the class of problems reviewed here, we consider "missing data" to be synonymous with "incomplete data." After describing common examples with missing data in Section 2, in Section 3 we describe techniques for handling these problems. In Section 4 we discuss the EM algorithm, an ubiquitous algorithm for finding maximum likelihood (m.l.) estimates from incomplete data. Useful reviews of the analysis of incomplete data, are given in Afifi and Elashoff (1966), Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird and Rubin (1977), and Little (1982).

2. Common Incomplete Data Problems

We first consider problems where missing values are confined to a single outcome variable y, and interest concerns the distribution of y, perhaps conditional on a set of one or more predictor variables x, that are recorded for all units in the sample. Sometimes we have no information about the missing values of y; at other times we may have partial information, for example, that they lie beyond a known censoring point c.

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.

2. Mechanisms Leading to Missing Values

Any analysis of incomplete data requires certain assumptions about the distribution of the missing values, and in particular how the distributions of the missing and observed values of a variable are related. The work of Rubin (1976a) distinguishes three cases. If the process leading to missing y values (and in particular, the probability that a particular value of y is missing) does not depend on the values of x or y, then the missing data are called missing at random and the observed data are observed at random. If the process depends on observed values of x and y but not on missing values of y the missing data are called missing at random, but the observed data are not observed at random. If the process depends on missing values of y then the missing data are not missing at random; in this case, particular care is required in deriving inferences. Rubin (1976a) formalizes these notions by defining a random variable m that indicates for each unit whether y is observed or missing, and relating these conditions to properties of the conditional distribution of m given x and y.

2.2 Analysis of Variance

The first incomplete data problem to receive systematic attention in the statistics literature is that of missing data in designed experiments; in the context of agricultural trials, this problem is often called the missing plot problem (Bartlett, 1937; Anderson, 1946). Designed experiments investigate the dependence of an outcome variable, such as yield of a crop, on a set of factors, such as variety, type of fertilizer and temperature. Usually an experimental design is chosen that allows efficient estimation of important effects as well as a simple analysis. The analysis is especially simple when the design matrix is easily inverted, as with complete or fractional replications of factorial designs. The missing data problem arises when at

the conclusion of the experiment, the values of the outcome variable are missing for some of the plots, perhaps because no values were possible, as when particular plots were not amenable to seeding, or because values were recorded and then lost. Standard analyses of the resultant incomplete data assume the missing data are missing at random, although in practical situations the plausibility of this assumption needs to be checked. The analysis aims to exploit the "near-balance" of the resulting data set to simplify computations. For example, one tactic is to substitute estimates of the missing outcome values and then to carry out the analysis assuming the data to be complete. Questions needing attention then address the choice of appropriate values to substitute and how to modify subsequent analyses to allow for such substitutions. For discussions of this and other approaches, see Healy and Westmacott (1956), Wilkinson (1958), and Rubin (1972, 1976b).

2.3 Censored or Truncated Outcome Variable

We have noted that standard analyses for missing plots assume that the missing data are missing at random, that is, the probability that a value is missing can depend on the values of the factors but not on the missing outcome values. This assumption is violated, for example, when the outcome variable measures time to an event (such as death of an experimental animal, failure of a light bulb), and the times for some units are not recorded because the experiment was terminated before the event had occurred; the resulting data are censored. In such cases the analysis must include the information that the units with missing data are censored, since if these units are simply discarded, the resulting estimates can be badly biased.

The analysis of censored samples from the Poisson, binomial and negative binomial distributions is considered by Hartley (1958). Other distributions, including the normal, log-normal, exponential, gamma, Weibull, extreme value

and logistic are covered most extensively in the life testing literature (for reviews, see Mann, Schafer and Singpurwalla, 1974; Tsokos and Shimi, 1977). Non-parametric estimation of a distribution subject to censoring is carried out by life table methods, formal properties of which are discussed by Kaplan and Meier (1958). Much of this work can be extended to handle covariate information (Glasser, 1969; Cox, 1972; Aitkin and Clayton, 1980; Laird and Oliver, 1981). The EM algorithm, discussed here in Section 4, is a useful computational device for such problems.

A variant of censored values occurs when missing values are known to lie within an interval, as when the data are available in grouped form. The analysis of grouped data is discussed by Hartley (1958), Kuldorff (1961) and Blight (1970), among others. Another variant of censored data occurs when the number of censored values is unknown. The resulting data are called truncated, since they can be regarded as a sample from a truncated distribution. A considerable literature exists for this form of data (Hartley, 1958; Dempster, Laird and Rubin, 1977; Blumenthal, Dahiya and Gross, 1978).

2.4 Sample Survey Data

For the data types discussed in Section 2.4, the missing data are not missing at random, but the mechanisms leading to incomplete data are assumed known. For example, the censoring points for censored observations are known. A common and somewhat more intractable problem occurs when the missing data are not missing at random and the mechanism leading to missing data is at best partially known. Incomplete data arising from nonresponse in sample surveys provide an illustration of this kind of problem. For example, nonresponse to a question on household income often depends on the amount of that income, in an unknown way. Restricting the analysis to respondents

clearly leads to bias in such situations; given the large samples often available in survey work, this bias is frequently more important than the loss of efficiency of estimation arising from the reduction in sample size.

The effect of survey nonresponse is minimized by: (a) designing data collection methods to minimize the level of nonresponse, (b) interviewing a subsample of nonrespondents, and (c) collecting auxilliary information on nonrespondents and employing analytical methods that use this information to reduce nonresponse bias. Models for nonrandomly missing data, as developed by Nelson (1976), Heckman (1976) and Rubin (1977), can also be applied here. Estimates derived from these models, however, are sensitive to aspects of the model that cannot be tested with the available data (Rubin, 1978; Little, 1982; Greenlees, Reece, and Zieschang, 1982). A thorough discussion of survey nonresponse is given in the work of the National Academy of the Sciences Panel on Incomplete Data (National Academy of the Sciences, 1982).

2.5 Multivariate Incomplete Data

The incomplete data structures discussed so far are univariate, in the sense that the missing values are confined to a single outcome variable. We now turn to incomplete data structures that are essentially multivariate in nature.

Many multivariate statistical analyses including least squares regression, factor analysis and discriminant analysis are based on an initial reduction of the data to the sample mean vector and covariance matrix of the variables. The question of how to estimate these moments with missing values in one or more of the variables is, therefore, an important one. Early literature was concerned with small numbers of variables (two or three) and simple patterns of missing data (Anderson, 1957; Afifi and Elashoff, 1966).

Subsequently, more extensive data sets with general patterns of missing data were addressed (Buck, 1960; Orchard and Woodbury, 1972; Trawinski and Bargmann, 1972; Rubin, 1974; Beale and Little, 1975; Little, 1976).

The reduction to first and second moments is generally not appropriate when the variables are categorical. In this case, the data can be expressed in the form of a multiway contingency table. Most of the work on incomplete contingency tables has concerned maximum likelihood estimation assuming a Poisson or multinomial distribution for the cell counts. Bivariate categorical data form a two-way contingency table; if some observations are available on a single variable only, then they can be displayed as a supplemental margin. The analysis of data with supplemental margins is discussed by Hocking and Oxspring (1974) and Chen and Fienberg (1974). Extensions to log-linear models for higher way tables with supplemental margins are discussed in Fuchs (1982).

Essentially, all of the literature on multivariate incomplete data assumes that the missing data are missing at random, and much of it also assumes that the observed data are observed at random. Together these assumptions imply that the process that creates missing data does not depend on any values, missing or observed.

3. Methods for Handling Incomplete Data

3.1 A Broad Taxonomy of Methods

Methods for handling incomplete data generally belong to one or more of the following categories:

(i) Methods that discard units with data missing in some variables and analyze only the units with complete data (for example, Nie et al, 1975).

(ii) Imputation based procedures. The missing values are filled in and the resultant completed data are analyzed by standard methods. For valid inferences to result, modifications to the standard analyses are required to allow for the differing status of the real and the imputed values. Commonly used procedures for imputation include hot deck imputation (c.f., Ford, 1981), where recorded units in the sample are substituted, mean imputation, where means from sets of recorded values are substituted and regression imputation, where the missing variables for a unit are estimated by predicted values from regression on the known variables for that unit (Buck, 1960). A variant of imputation methods produces multiple imputations for each missing value and thereby allows simple adjustments to be made to reflect the differing status of real and imputed values (Rubin, 1978, 1980).

(iii) Weighting procedures. Randomization inferences from sample survey data without nonresponse are commonly based on design weights, which are inversely proportional to the probability of selection. For example, let y_i be the value of a variable y for unit i in the population. Then, the population mean is often estimated by

$$\Sigma\ \pi_i^{-1} y_i \ /\ \Sigma\ \pi_i^{-1} \qquad (1)$$

where the sums are over sampled units, π_i is the probability of selection for unit i and π_i^{-1} is the design weight for unit i.

Weighting procedures modify the weights to allow for nonresponse. The estimator (1) is replaced by

$$\Sigma(\pi_i \hat{p}_i)^{-1} y_i \ /\ \Sigma(\pi_i \hat{p}_i)^{-1} \ , \qquad (2)$$

where the sums are now over sampled units which respond, and $\hat{p}_i$

is an estimate of the probability of response for unit i, usually the proportion of responding units in a subclass of the sample. Weighting is related to mean imputation; for example, if the design weights are constant in subclasses of the sample, then imputing the subclass mean for missing units in each subclass, or weighting responding units by the proportion responding in each subclass, lead to the same estimates of population means, although not the same estimates of sampling variance unless adjustments are made to the data with means imputed. A recent discussion of weighting with extensions to two way classifications is provided by Scheuren (1982).

(iv) Model-based procedures. A broad class of procedures is generated by defining a model for the incomplete data and basing inferences on the likelihood under that model, with parameters estimated by procedures such as maximum likelihood. Advantages of this approach are: flexibility; the avoidance of adhocery, in that model assumptions underlying the resulting methods can be displayed and evaluated; and the availability of large sample estimates of variance based on second derivatives of the log-likelihood, which take into account incompleteness in the data. Disadvantages are that computational demands can be large, particularly for complex patterns of missing data, and that little is known about the small sample properties of many of the large sample approximations.

3.2 The Modelling Approach to Incomplete Data

Any procedure that attempts to handle incomplete data must, either implicitly or explicitly, model the process that creates missing data. We prefer the explicit approach since assumptions are then clearly stated.

The parametric form of the modelling argument can be expressed as follows (Rubin, 1976a). Let y_p denote data that are present and y_m data that are missing. Suppose that $y = (y_p, y_m)$ has a distribution $f(y_p, y_m|\theta)$ indexed by an unknown parameter θ. If the missing data are missing at random, then the likelihood of θ given data y_p is proportional to the density of y_p, obtained by integrating $f(y_p, y_m|\theta)$ over y_m:

$$L(\theta|y_p) \propto \int f(y_p, y_m|\theta)dy_m \quad . \tag{3}$$

Likelihood inferences are based on $L(\theta|y_p)$. Occasionally in the literature, the missing values y_m are treated as fixed parameters, rather than integrated out of the distribution $f(y_p, y_m|\theta)$, and joint estimates of θ and y_m are obtained by maximizing $f(y_p, y_m; \theta)$ with respect to θ and y_m (e.g. Press and Scott, 1976 present a procedure which is essentially equivalent to this). This approach is not recommended since it can produce badly biased estimates which are not even consistent unless the fraction of missing data tends to zero as the sample size increases. Also, the model relating the missing and observed values of y is not fully exploited, and if the amount of missing data is substantial, the treatment of y_m as a set of parameters contradicts the general statistical principle of parsimony.

An important generalization of (3) is to include in the model the distribution of a vector of variables indicating whether a value is observed or missing. The full distribution can be specified as

$$f(m, y_p, y_m|\theta, \phi) = f(y_p, y_m|\theta)f(m|y_p, y_m, \phi) \quad , \tag{4}$$

where θ is the parameter of interest and ϕ relates to the mechanism leading to missing data. This extended formulation is necessary for nonrandomly missing data such as arise in censoring problems.

To illustrate (3) and (4), suppose the hypothetical complete data $y = (y_1,\ldots,y_n)$ is a random sample of size n from the exponential distribution with mean θ. Then

$$f(y_p,y_m|\theta) = \theta^{-n} \exp(-t_n/\theta) \quad ,$$

where $t_n = \sum_{i=1}^{n} y_i$ is the total of the n sampled observations. If $r < n$ observations are present and the remaining $n-r$ are missing, then the likelihood ignoring the response mechanism is proportional to the density

$$f(y_p|\theta) = \theta^{-r} \exp(-t_r/\theta) \quad , \tag{5}$$

regarded as a function of θ, where t_r is the total of the recorded observations.

Let $m = (m_1,\ldots,m_n)$ where $m_i = 1$ or 0 as y_i is recorded or missing, respectively, $r = \Sigma\, m_i$. We consider two models for the distribution of m given y. First, suppose observations are independently recorded or missing with probability ϕ. Then

$$f(m|y,\phi) = \phi^r(1-\phi)^{n-r} \quad ,$$

and

$$f(y_p,m|\theta,\phi) = \phi^r(1-\phi)^{n-r}\theta^{-r} \exp(-t_r/\theta) \quad . \tag{6}$$

The likelihoods based on (5) and (6) differ by a factor $\phi^r(1-\phi)^{n-r}$ which does not depend on θ, provided that θ and ϕ are distinct, that is their joint parameter space factorizes into a θ-space and a ϕ-space. Hence we can base inferences on (5), ignoring the response mechanism.

Suppose instead that the sample is censored, in that only values less than a known censoring point c are observed. Then

$$f(m|y,\phi) = \prod_{i=1}^{n} f(m_i|y_i) \quad ,$$

$$f(m_i|y_i) = \begin{cases} 1 & \text{if } m_i = 1 \text{ and } y_i < c \text{ or } m_i = 0 \text{ and } y_i < c \ ; \\ 0 & \text{otherwise} \ . \end{cases}$$

The full likelihood is then proportional to

$$\begin{aligned} f(y_p,m|\theta) &= \sum_{i:m_i=1} f(y_i|\theta) f(m_i|y_i < c) \sum_{i:m_i=0} pr(y_i > c|\theta) \\ &= \theta^{-r} \exp(-t_r/\theta) \exp[-(n-r)c/\theta] \ . \end{aligned} \tag{7}$$

In this case the response mechanism is not ignorable, and the likelihoods based on (5) and (7) differ. In particular, the maximum likelihood estimate of θ based on (5) is t_r/r, the mean of the recorded observations, which is less than the correct maximum likelihood estimate of θ based on (7), namely $[t_r+(n-r)c]/r$. The latter estimate has the simple interpretation as the total time at risk for the uncensored and censored observations divided by the number of failures (r).

3.3 Special Data Patterns: Factoring the Likelihood

For certain special patterns of multivariate missing data, maximum likelihood estimation can be simplified by factoring the joint distribution in a way which simplifies the likelihood. Suppose for example the data have the monotone or nested pattern in Figure 1, where y_j represents a set of variables observed for the same set of observations and y_j is more observed than y_{j+1}, $j = 1...,J-1$. The joint distribution of $y_1 \cdots y_J$ can be factored in the form

$$f(y_1,\ldots,y_J|\theta) = f_1(y_1|\theta_1)f_2(y_2|y_1,\theta_2) \cdots f(y_J|y_1,\ldots,y_{J-1},\theta_J) \ ,$$

where f_j denotes the conditional distribution of y_j given $y_1 \cdots, y_{j-1}$, indexed by parameters θ_j. If the parameters $\theta_1,\ldots\theta_J$ are distinct, then the likelihood of the data factors into distinct complete-data components, leading to simple maximum likelihood estimators for θ (Anderson, 1957; Rubin, 1974). Maximum likelihood estimation with more general patterns of incomplete data can be accomplished by the EM algorithm.

4. General Data Patterns: The EM Algorithm

The expectation-maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) is an iterative method of maximum likelihood estimation that applies to any pattern of missing data. Let $\ell(\theta|y_p,y_m)$ denote the log-likelihood of parameters θ based on the hypothetical complete data (y_p,y_m). Let $\theta^{(i)}$ denote an estimate of θ after iteration i of the algorithm. The $(i + 1)$th iteration consists of an E-step and an M-step. The E-step consists of taking the expectation of $\ell(\theta|y_p,y_m)$ over the conditional distribution of y_m given y_p, evaluated at $\theta = \theta^{(i)}$. That is, the averaged loglikelihood

$$\ell^*(\theta|y_p,\theta^{(i)}) = \int \ell(\theta|y_p,y_m)f(y_m|y_p,\theta^{(i)})dy_m$$

is formed.

The M-step consists in finding $\theta^{(i+1)}$, the value of θ which maximizes ℓ^*. This new estimate, $\theta^{(i+1)}$, then replaces $\theta^{(i)}$ at the next iteration. Each step of EM increases the loglikelihood of θ given y_p, $\ell(\theta|y_p)$. Under quite general conditions, the algorithm converges to a maximum value of the loglikelihood $\ell(\theta|y_p)$. In particular, if a unique finite maximum likelihood estimate of θ exists, the algorithm finds it.

An important case occurs when the complete data belong to a regular exponential family. In this case, the E-step reduces to estimating the sufficient statistics corresponding to the natural parameters of the distribution. The M-step corresponds to maximum likelihood estimation from the hypothetical complete data, with the sufficient statistics replaced by the estimated sufficient statistics from the E-step.

The EM algorithm was first introduced for particular problems (e.g., Hartley, 1958, for counted data and Blight, 1970, for grouped or censored data). The regular exponential family case was presented by Sundberg (1974). Orchard and Woodbury (1972) discussed the algorithm more generally, using the term "missing information principle" to describe the link with the complete-data loglikelihood. Dempster, Laird and Rubin (1977) introduced the term EM, developed convergence properties and provided a large body of examples. Recent applications include missing data in discriminant analysis (Little, 1978) and regression with grouped or censored data (Hasselblad, Stead, and Galke, 1980).

The EM algorithm converges reliably, but it has slow convergence properties if the amount of information in the missing data is relatively large. Also, unlike methods like Newton-Raphson that need to calculate and invert an information matrix, EM does not provide asymptotic standard errors for the maximum likelihood estimates as output from the calculations. Its

popularity derives from its link with maximum likelihood for complete data and its consequent usually simple computational form. The M-step often corresponds to a standard method of analysis for complete data and thus can be carried out with existing technology. The E-step often corresponds to imputing values for the missing data y_m, or more generally, for the sufficient statistics that are functions of y_m and y_p, and as such relates maximum likelihood procedures to imputation methods. For example, the EM algorithm for multivariate normal data can be viewed as an iterative version of Buck's (1960) method for imputing missing values (Beale and Little, 1975).

Although the EM algorithm is a powerful tool for estimation from incomplete-data, many problems remain. For example, nonnormal likelihoods occur more commonly with incomplete data than with complete data, and much remains to be learned about the appropriateness of many incomplete-data methods when applied to real data.

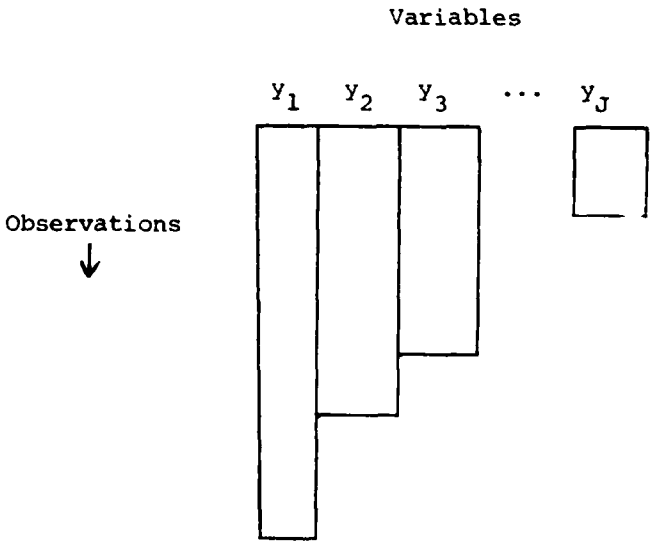


Figure 1. Schematic representation of a monotone (or nested) data pattern

REFERENCES

Afifi, A. A. and Elashoff, R. M. "Missing observations in multivariate statistics I: Review of the literature." J. Am. Statist. Assoc., 61 (1966), pp. 595-604.

Aitkin, M. and Clayton, D. (1980). The Fitting of Exponential, Weibull and Extreme Value Distributions to Complex Censored Survival Data Using GLIM. Applied Statistics, 156-163.

Anderson, R. L. "Missing plot techniques." Biometrics, 2 (1946), pp. 41-47.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Am. Statist. Ass. 52, 200-3.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. J. Roy. Statist. Soc. B 37, 129-146.

Blight, B. J. N. (1970). Estimation from a censored sample for the exponential family. Biometrika, 57, 389-395.

Blumenthal, S., Dahiya, R. C. and Gross, A. S. (1978). Estimating the Complete Sample Size from an Incomplete Poisson Sample. J. Am. Statist. Assoc. 73, 182-187.

Buck, S. F. (1960). "A method of estimation of missing values in multivariate data, suitable for use with an electronic computer," J. Roy. Statist. Soc., B, 22, pp. 302-306.

Chen, T. and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially classified data. Biometrics 30, 629-642.

Cox, D. R. (1972). Regression models and life tables (with discussion). J. Roy. Statist. Soc., B, 34, 187-220.

Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. Ann. Math. Statist. 43, 1470-1480.

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

Ford, B. N. (1982). An overview of hot deck procedures. In Nonresponse in Sample Surveys: Theory of Current Practice. Academic Press (in press).

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. J. Am. Statist. Assoc. 77.

Glasser, M. (1967). Exponential Survival with Covariance. J. Am. Statist. Assoc. 62, 561-568.

Greenlees, W. S., Reece, J. S. and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends upon the variable being imputed. J. Amer. Statist. Assoc. 77.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. Biometrics 14, 174-194.

Hartley, H. O. and Hocking, R. R. (1971). The analysis of incomplete data. Biometrics 27, 783-808.

Hasselblad, V., Stead, A. G. and Galke, W. (1980). Application of Regression Analysis to Grouped Blood Lead Data. J. Am. Statist. Assoc., 75, 771-779.

Healy, M. and Westmacott, M. (1956). Missing values in experiments analysed on automatic computers. Appl. Statist. 5, 203-206.

Heckman, J. D. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475-492.

Hocking, R. R. and Oxspring, H. H. (1974). The analysis of partially categorized contingency data. Biometrics 30, 469-483.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. J. Roy. Statist. Assoc., 53, 457-481.

Kulldorff, G. (1961). Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples. Almquist and Wiksell, Stockholm and Wiley, New York.

Laird, N. and Olivier, D. (1981). Covariance analysis of survival data using log-linear analysis techniques. J. Am. Statist. Assoc. 76, 231-240.

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

Little, R. J. A. (1978). Consistent Methods for Discriminant Analysis with Incomplete Data. J. Am. Statist. Assoc., 73, 319-322.

Little, R. J. A. (1982). Models for Nonresponse in sample surveys. J. Am. Statist. Assoc. 77.

Mann, N. R., Schafer, R. E. and Singpurwalla, N. D. Methods for Statistical Analysis of Reliability and Life Data. John Wiley & Sons, New York.

National Academy of Sciences (1982). Report of the Panel on Incomplete Data. National Academy of Sciences, Washington, DC.

Nelson, F. D. (1977). Censored regression models with unobserved, stochastic censoring thresholds. Journal of Econometrics 6, 581-92.

Nie, N. H. Hull, C. H., Jenkins, J. G., Steinbrenner, K. and Bent, D. H. (1975). SPSS, Second Edition, McGraw Hill.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. Proc. 6th Berkeley Symposium on Math. Statist. and Prob. 1, 697-715.

Press, S. J. and Scott, A. J. (1976). Missing Variables in Bayesian Regression II. J. Am. Statist. Assoc., 71, 366-369.

Rubin, D. B. (1972). A noniterative algorithm for least squares estimation of missing values in any analysis of variance design. Appl. Statist., 21, 136-141.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete data problems. J. Am. Statist. Ass., 69, 467-474.

Rubin, D. B. (1976a). Inference and missing data. Biometrika, 63, 581-92.

Rubin, D. B. (1976b). Noniterative least squares estimates, standard errors, and F-tests for any analysis of variance design with missing data. J. Roy. Statist. Soc. B, 38, 3, 270-274.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. J. Am. Statist. Assoc., 72, 538-543.

Rubin, D. B. (1978). "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse" (with discussion and reply), in Imputation and Editing of Faulty or Missing Survey Data, U. S. Social Security Administration and Bureau of the Census, 1-9.

Rubin, D. B. (1980). Handling nonresponse in sample surveys by multiple imputations. U. S. Department of Commerce, Bureau of the Census Monograph.

Scheuren, F. (1982). Weighting Adjustment for Unit Nonresponse. Nonresponse in Sample Surveys: Theory of Current Practice. Academic Press (in press).

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scand. J. Statist., 1, 49-58.

Trawinski, I. M. and Bargmann, R. E. (1964). "Maximum likelihood estimates with incomplete multivariate data." Ann. Math. Statist., 35, pp. 647-657.

Tsokos, C. P. and Shimi, I. N. (1977). The thecry and applications of Reliability. Academic Press, New York.

Wilkinson, G. N. (1958). "The analysis of variance and derivation of standard errors for incomplete data," Biometrics, 14, pp. 360-384.

RJAL/DBR/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER #2372	2. GOVT ACCESSION NO. AD A116188	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Incomplete Data - Encyclopedia Entry		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Roderick J. A. Little and Donald B. Rubin		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 3 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE April 1982 13. NUMBER OF PAGES 19
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Missing Data, EM Algorithm, Nonresponse, Censored Data, Truncated Data, Imputation, Factorizing Likelihoods

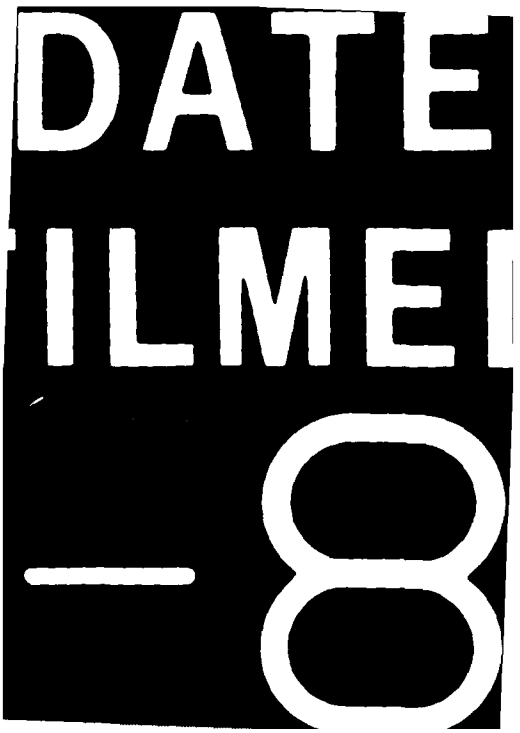
20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Incomplete data occur commonly in the application of statistics to real data, and there exists a substantial literature on the problem. This article provides an overview of issues for the statistically knowledgeable but not necessarily statistically sophisticated reader. It is intended to appear as an entry in The Encyclopedia of Statistical Sciences.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)


DATE
ILME
8