DOCUMENT RESUME

ED 075 486 TM 002 565

AUTHOR	Rubin, Donald B.
TITLE	Missing at Random: What Does it Mean? Draft.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-73-2
PUB DATE	Jan 73
NOTE	11p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bulletins; *Data Analysis; *Mathematical Models; *Statistical Analysis; Technical Reports

ABSTRACT

Most articles on missing values assume the missing data are "missing at random" and ignore the process that "caused" the missing values. The condition under which this procedure is justified is explored here: the concept of missing at random is precisely defined, several examples are discussed, and two simple conditions are given which are sufficient to assure that the missing data are missing at random. (Author)

RB-73-2

RESEARCH BULLETIN

MISSING AT RANDOM - WHAT DOES IT MEAN?

Donald B. Rubin

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
January 1973

MISSING AT RANDOM - WHAT DOES IT MEAN?

Donald B. Rubin
Educational Testing Service

Abstract

Most articles on missing values assume the missing data are "missing at random" and ignore the process that "caused" the missing values. The condition under which this procedure is justified is explored here: the concept of missing at random is precisely defined, several examples are discussed, and two simple conditions are given which are sufficient to assure that the missing data are missing at random.

MISSING AT RANDOM - WHAT DOES IT MEAN?

Donald B. Rubin
Educational Testing Service

1. "Missing at Random" as Used in the Literature

In many articles on missing values there is an assumption either implicit or explicit that the missing data are "missing at random" in the sense that the process that caused the missing values can be ignored. In some articles such as those concerned with the multivariate normal (Afifi & Elashoff, 1966; Anderson, 1957; Hartley & Hocking, 1971; Hocking & Smith, 1968; Wilks, 1932), "missing at random" seems to mean that each item in the data matrix is equally likely to be missing. In other articles such as those dealing with the analysis of variance (Hartley, 1956; Healy & Westmacott, 1956; Rubin, 1972; Wilkinson, 1958), "missing at random" seems to mean that observations of the dependent variable are missing without regard to the actual values that would have been observed. Similarly, "missing at random" apparently can mean missing according to a preplanned experimental design (Hocking & Smith, 1972; Trawinski & Bargmann, 1964).

The objective here is to explore the specific assumptions that need to be made in order to ignore the process that caused the missing values when investigating the density of the data. More specifically, the approach will be to examine the likelihood function of the observed data and the observed pattern of missing values and then to specify the condition under which solutions (e.g., maximum likelihood estimates and sampling distributions, Bayes posterior distributions) based on this likelihood agree with those based on the marginal likelihood of the observed data.

2. Notation and a Definition

Let P_θ^V be a probability density function for a real-valued vector random variable V of order k, where θ is a vector parameter which lies in an open parameter space Ω. A sample realization of V is the data. Generally $k = pn$ where p = number of "variables" and n = number of "units." We assume that the data analyst's primary objective is the investigation of this density (e.g., estimating θ, testing hypotheses about θ, estimating a posterior density for θ). Let W be a 0-1 indicator random variable of length k, and let $P_\xi^{V,W}$ be the joint probability density function for V and W where $\xi\epsilon\Omega$ is the vector parameter for this density. A sample realization of W will indicate the missing values in the data. We have, of course, that $P_\theta^V = \int_W P_\xi^{V,W}$ where $\int_W$ is the integral over the W random variable. We also define $P_\phi^{W\cdot V} = P_\xi^{V,W}/P_\theta^V$ to be the conditional density of the missing value indicator given the data where $\phi\epsilon\Omega$.

Let v,w be a sample realization of V,W. If $w_i = 1$, v_i is an observed scalar random variable and thus is a real number. If $w_i = 0$, v_i is an unobserved scalar random variable. Thus v is composed of k-m real numbers and m unobserved scalar random variables, where m is the number of missing values. Let $\overset{\circ}{v}$ indicate the m-vector of unobserved random variables in v, i.e., the missing data.

The likelihood function of all observables, that is, the indicator variable and the observed data, is

$$(1) \qquad \int_{\overset{\circ}{v}} P_\xi^{V,W}(v,w)$$

where $P^{V,W}_{\xi}$ (v,w) is the density of V,W evaluated at the observed values of w and v regarded as a function of $\overset{\circ}{v}$ and the parameters ξ, and $\int_{\overset{\circ}{v}}$ represents the integral over $\overset{\circ}{v}$, the unobserved scalar random variables. This likelihood can also be written as

$$\left[\int_{\overset{\circ}{v}} P^{V}_{\theta}\ (v)\right]\left[\int_{\overset{\circ}{v}} P^{V,W}_{\xi}\ (v,w)\ /\int_{\overset{\circ}{v}} P^{V}_{\theta}\ (v)\right]$$

where

$$(2) \qquad \int_{\overset{\circ}{v}} P^{V}_{\theta}\ (v)$$

is the marginal likelihood of the observed data and

$$(3) \qquad \int_{\overset{\circ}{v}} P^{V,W}_{\xi}\ (v,w)\ /\int_{\overset{\circ}{v}} P^{V}_{\theta}\ (v)$$

is the conditional likelihood of the missing value indicator given the observed data.

Definition: The missing data $\overset{\circ}{v}$ are said to be missing at random if the conditional likelihood of the missing value indicator given the observed data, equation (3), is independent of θ.

The motivation for this definition is that when the data are missing at random, maximum likelihood estimates of θ and their sampling distributions (as well as Bayes posterior densities for θ) obtained from the marginal likelihood of the observed data, equation (2), agree with those obtained from the full likelihood of all observables, equation (1). In this sense, if the data are missing at random, the observed data may be said to be "sufficient" for P^{V}_{θ}.

3. Two Simple Conditions Sufficient for the Missing Data to be Missing at Random

By rewriting $P_{\xi}^{V,W}(v,w)$ as $P_{\theta}^{V}(v)\, P_{\phi}^{W\cdot V}(v,w)$ we have that equation (3). the conditional likelihood of the missing value indicator given the observed data, can be written as

$$(4) \qquad \int_{\overset{\circ}{v}} P_{\theta}^{V}(v)\, P_{\phi}^{W\cdot V}(v,w) \Big/ \int_{\overset{\circ}{v}} P_{\theta}^{V}(v) \quad .$$

Clearly, if $P_{\phi}^{W\cdot V}(v,w)$ is independent of θ and $\overset{\circ}{v}$, equation (4) is independent of θ; hence, the following result.

Lemma: If (1) $P_{\phi}^{W\cdot V}(v,w)$ is independent of $\overset{\circ}{v}$, the missing data, and

(2) θ and ϕ lie in disjoint parameter spaces,

then the missing data, $\overset{\circ}{v}$, are missing at random.

The first condition in this lemma is satisfied by all of the examples given by the references cited in Section 1. "Equally likely" missing values in the data matrix yield

$$P_{\phi}^{W\cdot V}(w,v) = \prod_{i=1}^{k} \phi^{w_i}(1-\phi)^{1-w_i} \quad , \qquad w_i = 0 \text{ or } 1 \quad ,$$

where ϕ is the probability of being observed. "Preplanned" missing observations yield

$$P_{\phi}^{W\cdot V}(w,v) = \prod_{i=1}^{k} \delta(w_i - \phi_i) \quad ; \qquad w_i = 0 \text{ or } 1 \quad ,$$

where ϕ is the 0-1 vector indicating the preplanned pattern of missing observations and $\delta(a) = 1$ if $a = 0$ and zero otherwise. "Without regard to values that would have been observed" simply implies that $P_{\phi}^{W\cdot V}(w,v)$ is independent of the missing values $\overset{\circ}{v}$. As a more complex example, assume that odd v_i

are always observed and even v_i are missing if the preceding value v_{i-1} is greater than Φ. Letting $T_1 = \{\text{odd } i, i = 1,\ldots,k\}$ and $T_2 = \{\text{even } i, i = 2,\ldots,k\}$ we have

$$P_\Phi^{W \cdot V} = \prod_{i \in T_1} \delta(1 - w_i) \prod_{i \in T_2} \gamma(w_i, v_{i-1} - \Phi)$$

$$\gamma(a,b) = \begin{cases} 1 & \text{if } \begin{cases} a = 0 \text{ and } b \geq 0 \text{ , or} \\ a = 1 \text{ and } b < 0 \end{cases} \\ 0 & \text{otherwise} \end{cases} .$$

If in these examples θ and Φ lie in disjoint parameter spaces, both conditions in the lemma are satisfied and the missing data will be missing at random. If condition (1) in the lemma is satisfied but condition (2) is not, it is clear from equation (4) that the data are <u>not</u> missing at random; nevertheless, maximum likelihood and Bayes procedures applied to the marginal likelihood of the observed data $\int_{\overset{o}{v}} P_\theta^V (v)$ are "reasonable" (e.g., consistent) and suffer only from reduced "efficiency". Thus, in a sense, condition (1) in the lemma might have been chosen as the definition of missing at random. However, then discussion of maximum likelihood and Bayes procedures following an assumption of missing at random would always be somewhat imprecise and inaccurate.

An argument could be made for choosing conditions (1) and (2) of the lemma as the definition of missing at random because mcdels not satisfying condition (1) intuitively do not seem to have missing data missing at random. For example, assume the data for odd i are uniform on $(0,\theta)$, and the data for even i are uniform on $(0,1)$ and missing if less than Φ (θ and Φ lie in disjoint parameter spaces); then by equation (3) the data are missing at random even though condition (1) is not satisfied.

Nevertheless, if the phrase "missing at random" is meant to imply that the process that caused the missing values, whatever it may be, can be ignored, the definition of missing at random given here in Section 3 is appropriate.

4. Examples

As a practical missing values problem consider the problem of nonresponse in sample surveys, where the parameters θ are the parameters of the joint distribution of response variables and background variables. Assume the nonrespondents are known to be typically different from the respondents, say, to have lower socioeconomic status (SES). Are the data missing at random? Assume the researcher has recorded a measure of SES as well as other potentially relevant background variables for all subjects. If conditionally given these observed background variables, a subject will offer or not offer his response independently of what that response would be, that is, if subjects with identical background variables (but possibly different responses) are equally likely to respond, then condition (1) in the lemma is satisfied; if, in addition, the parameters of the nonresponse process are independent of θ, the missing data are missing at random. Hence by collecting "additional" variables the researcher can often make the assumption of missing at random plausible.

However, even if the missing data are missing at random, the researcher's problem in choosing an appropriate model may be more serious than it would be if there were no missing data. For example, if the regression of response variables on background variables is curvilinear, and there are many missing responses when the values of the background variables are extreme (e.g., low SES), fitting a linear model may yield especially poor prediction of the typical responses for those subjects likely to have missing responses.

As another example of missing data consider nonresponse on multiple choice questionnaires. Lord (1973) makes the distinction between "not reached" items, which the examinee did not have time to attempt, and "omitted" items, which the examinee reached, presumably read, but did not answer. θ includes the parameters of subject ability and item difficulty. If the items on the test are not ordered with respect to difficulty, it seems reasonable to assume, as does Lord, that condition (1) in the lemma holds for the not-reached items but does not hold for the omitted items; that is, $P_{\phi}^{W \cdot V}(w,v)$ is independent of the $\overset{\circ}{v}$ corresponding to the not-reached items but does depend upon the $\overset{\circ}{v}$ corresponding to the omitted items. However, it also seems fairly clear that the parameters ϕ and θ may not lie in disjoint parameter spaces since more intelligent examinees probably reach more items and omit a lower proportion of items reached. Assuming that the number of items reached does not depend upon θ, then the not-reached items are missing at random.

The investigation of complex models for nonrandom missing values such as might be appropriate for Lord's data set is a relatively unexplored area of statistics. Only a few "censored-data" models are commonly available for dealing with nonrandomly missing data (e.g., see Swan, 1969).

References

Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics. I: Review of the literature. *Journal of the American Statistical Association*, *61*, 595-604.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, *52*, 200-203.

Hartley, H. O. (1956). Programming analysis of variance for general purpose computers. *Biometrics*, *12*, 110-122.

Hartley, H. O., & Hocking, R. R. (1971). Incomplete data analysis. ENAR-IMS Spring Regional Presidential Invited Lecture. (To be published in *Biometrics*.)

Healy, M. J. R., & Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Applied Statistics*, *5*, 203-206.

Hocking, R. R., & Smith, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. *Journal of the American Statistical Association*, *63*, 159-173.

Hocking, R. R., & Smith, W. B. (1972). Optimum incomplete multinormal samples. *Technometrics*, *14*, 299-307.

Lord, F. M. (1973). Estimation of latent ability and item parameters when there are omitted responses. Educational Testing Service Research Bulletin.

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

Swan, A. V. (1969). Computing maximum likelihood estimates for parameters of the normal distribution from grouped and censored data. Applied Statistics, 18, 65-69.

Trawinski, I. M., & Bargmann, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. Annals of Mathematical Statistics, 35, 647-657.

Wilkinson, G. N. (1958). Estimation of missing values for the analysis of incomplete data. Biometrics, 14:2, 257-286.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics, 3, 163-195.