Biometrika (1976), **63**, 3, *pp.* 581–92
Printed in Great Britain

Inference and missing data

By DONALD B. RUBIN
Educational Testing Service, Princeton, New Jersey

Summary

When making sampling distribution inferences about the parameter of the data, θ, it is appropriate to ignore the process that causes missing data if the missing data are 'missing at random' and the observed data are 'observed at random', but these inferences are generally conditional on the observed pattern of missing data. When making direct-likelihood or Bayesian inferences about θ, it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is 'distinct' from θ. These conditions are the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences.

Some key words: Bayesian inference; Incomplete data; Likelihood inference; Missing at random; Missing data; Missing values; Observed at random; Sampling distribution inference.

1. Introduction: The generality of the problem of missing data

The problem of missing data arises frequently in practice. For example, consider a large survey of families conducted in 1967 with many socioeconomic variables recorded, and a follow-up survey of the same families in 1970. Not only is it likely that there will be a few missing values scattered throughout the data set, but also it is likely that there will be a large block of missing values in the 1970 data because many families studied in 1967 could not be located in 1970. Often, the analysis of data like these proceeds with an assumption, either implicit or explicit, that the process that caused the missing data can be ignored. The question to be answered here is: when is this the proper procedure?

The statistical literature on missing data does not answer this question in general. In most articles on unintended missing data, the process that causes missing data is ignored after being assumed accidental in one sense or another. In some articles such as those concerned with the multivariate normal (Afifi & Elashoff, 1966; Anderson, 1957; Hartley & Hocking, 1971; Hocking & Smith, 1968; Wilks, 1932), the assumption about the process that causes missing data seems to be that each value in the data set 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, 1976; Wilkinson, 1958), the assumption seems to be that values of the dependent variables are missing without regard to values that would have been observed.

The statistical literature also discusses missing data that arise intentionally. In these cases, the process that causes missing data is generally considered explicitly. Some examples of methods that intentionally create missing data are: a preplanned multivariate experimental design (Hocking & Smith, 1972; Trawinski & Bargmann, 1964); random sampling from a finite population, i.e. the values of variables for unsampled units being missing (Cochran, 1963, p. 18); randomization in an experiment, where, for each unit, the values that would have been observed had the unit received a different treatment are missing

(Kempthorne, 1952, p. 137; Rubin, 1975); sequential stopping rules, where the values after the last one observed are missing (Lehmann, 1959, p. 97), and even some 'robust analyses', where observed values are considered outliers and so discarded or made missing.

2. OBJECTIVES AND BROAD REVIEW

Our objective is to find the weakest simple conditions on the process that causes missing data such that it is always appropriate to ignore this process when making inferences about the distribution of the data. The conditions turn out to be rather intuitive as well as nonparametric in the sense that they are not tied to any particular distributional form. Thus they should prove helpful for deciding in practical problems if the process that causes missing data can be ignored.

Section 3 gives the notation for the random variables: θ is the parameter of the data, and ϕ is the parameter of the missing-data process, i.e. the parameter of the conditional distribution of the missing-data indicator given the data. Section 4 presents examples of processes that cause missing data.

Section 5 shows that when the process that causes missing data is ignored, the missing-data indicator random variable is simply fixed at its observed value. Whether this corresponds to proper conditioning depends on the method of inference and three conditions on the process that causes missing data. These conditions place no restrictions on the missing-data process for patterns of missing data other than the observed pattern. Their formal definitions correspond to the following statements.

The missing data are missing at random if for each possible value of the parameter ϕ, the conditional probability of the observed pattern of missing data, given the missing data and the value of the observed data, is the same for all possible values of the missing data.

The observed data are observed at random if for each possible value of the missing data and the parameter ϕ, the conditional probability of the observed pattern of missing data, given the missing data and the observed data, is the same for all possible values of the observed data.

The parameter ϕ is distinct from θ if there are no *a priori* ties, via parameter space restrictions or prior distributions, between ϕ and θ.

Sections 6, 7 and 8 use these definitions to prove that ignoring the process that causes missing data when making sampling distribution inferences about θ is appropriate if the missing data are missing at random and the observed data are observed at random, but the resulting inferences are generally conditional on the observed pattern of missing data. Further, ignoring the process that causes missing data when making direct-likelihood or Bayesian inferences about θ is appropriate if the missing data are missing at random and ϕ is distinct from θ.

Other results show that these conditions are the weakest simple and general conditions under which it is always appropriate to ignore the process that causes missing data. The reader not interested in the formal details should be able to skim §§ 3–8 and proceed to § 9.

Section 9 uses these results to highlight the distinctions between the sampling distribution and the likelihood-Bayesian approaches to the problem of missing data. Section 10 concludes the paper with the suggestion that in many practical problems, Bayesian and likelihood inferences are less sensitive than sampling distribution inferences to the process that causes missing data.

Throughout, measure-theoretic considerations about sets of probability zero are ignored.

3. Notation for the random variables

Let $U = (U_1, \ldots, U_n)$ be a vector random variable with probability density function f_θ. The objective is to make inferences about θ, the vector parameter of this density. Often in practice, the random variable U will be arranged in a 'units' by 'variables' matrix. Let $M = (M_1, \ldots, M_n)$ be the associated 'missing-data indicator' vector random variable, where each M_i takes the value 0 or 1. The probability that M takes the value $m = (m_1, \ldots, m_n)$ given that U takes the value $u = (u_1, \ldots, u_n)$ is $g_\phi(m|u)$, where ϕ is the nuisance vector parameter of the distribution.

The conditional distribution g_ϕ corresponds to 'the process that causes missing data': if $m_i = 1$, the value of the random variable U_i will be observed while if $m_i = 0$, the value of U_i will not be observed. More precisely, define the extended vector random variable $V = (V_1, \ldots, V_n)$ with range extended to include the special value $*$ for missing data: $v_i = u_i$ $(m_i = 1)$, and $v_i = *$ $(m_i = 0)$. The values of the random variable V are observed, not the random variable U, although it is desired to make inferences about the distribution of U.

4. Examples of processes that cause missing data

In order to clarify the notation in § 3 we give four examples.

Example 1. Suppose there are n samples of an alloy and on each we attempt to record some characteristic by an instrument that has a constant probability, ϕ, of failing to record the result for all possible samples. Then

$$g_\phi(m|u) = \prod_{i=1}^{n} \phi^{m_i}(1-\phi)^{1-m_i}.$$

Example 2. Let u_i be the value of blood pressure for the ith subject $(i = 1, \ldots, n)$ in a hospital survey. Suppose $v_i = *$ if u_i is less than ϕ, which equals the mean blood pressure in the population; i.e. we only record blood pressure for subjects whose blood pressures are greater than average. Then

$$g_\phi(m|u) = \prod_{i=1}^{n} \delta\{\gamma(u_i - \phi) - m_i\},$$

where $\gamma(a) = 1$ if $a \geqslant 0$ and 0 otherwise; $\delta(a) = 1$ if $a = 0$ and 0 otherwise.

Example 3. Observations are taken in sequence until a particular function of the observations is in a specified critical region C. Here n is essentially infinite and, for some n_1 which is a function of the observations, $v_i \neq *$ $(i \leqslant n_1)$, and $v_i = *$ $(i > n_1)$. Thus

$$g_\phi(m|u) = \prod_{i=1}^{n_1} \delta(1 - m_i) \prod_{i=n_1+1}^{n} \delta(m_i),$$

where n_1 is the minimum k such that the function $Q_k(u_1, \ldots, u_k) \in C$.

Example 4. Let $n = 2$. If $u_1 \geqslant 0$: with probability ϕ, $v_1 \neq *$ and $v_2 = *$; and with probability $1-\phi$, $v_1 \neq *$ and $v_2 \neq *$. If $u_1 < 0$: with probability ϕ, $v_1 \neq *$ and $v_2 = *$; and with probability $1-\phi$, $v_1 = *$ and $v_2 \neq *$. Thus

$$g_\phi(m|u) = \begin{cases} \phi & \text{if } m = (1, 0), \\ (1-\phi)\gamma(u_1) & \text{if } m = (1, 1), \\ (1-\phi)\{1-\gamma(u_1)\} & \text{if } m = (0, 1), \\ 0 & \text{if } m = (0, 0). \end{cases}$$

5. IGNORING THE PROCESS THAT CAUSES MISSING DATA

Let $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_n)$ be a particular sample realization of V, i.e. each $\tilde{v}_i$ is either a known number or a missing value, $*$. These observed values imply an observed value for the random variable M, $\tilde{m} = (\tilde{m}_1, \ldots, \tilde{m}_n)$, and imply observed values for some of the scalar random variables in U. That is, if $\tilde{v}_i$ is a number, then the observed value of M_i, $\tilde{m}_i$, is one, and the observed value of U_i, $\tilde{u}_i$, equals $\tilde{v}_i$; if $\tilde{v}_i = *$, then $\tilde{m}_i = 0$ and the value of U_i is not known; in special cases, knowing values in $\tilde{v}$ may imply observed values for some U_i with $\tilde{v}_i = *$, for example f_θ specifies $u_1 = u_2 + u_3$ and we observe $\tilde{v}_1 = *$, $\tilde{v}_2 = 3{\cdot}1$ and $\tilde{v}_3 = 5{\cdot}2$.

Table 1. *Classifying the examples in* § 4

Example	Missing data, missing at random	Observed data, observed at random	ϕ distinct from θ
1	Always MAR	Always OAR	Always distinct
2	MAR only if all $\tilde{m}_i = 1$	OAR only if all $\tilde{m}_i = 0$	Distinct only if mean blood pressure in the population is known *a priori*
3	Always MAR	Never OAR	Always distinct
4	MAR unless $\tilde{m} = (0, 1)$	OAR unless $\tilde{m} = (1, 1)$	Distinct if *a priori* ϕ is not restricted by θ

Hence, the observed value of M, namely $\tilde{m}$, effects a partition of each of the vectors of random variables and the vectors of observed values into two vectors corresponding to $\tilde{m}_i = 0$ for missing data and $\tilde{m}_i = 1$ for observed data. For convenience write

$$U = (U_{(0)}, U_{(1)}), \quad V = (V_{(0)}, V_{(1)}), \quad u = (u_{(0)}, u_{(1)}), \quad v = (v_{(0)}, v_{(1)}),$$

where by definition $v_{(0)} = (*, \ldots, *)$ and $u_{(1)} = v_{(1)}$. It is important to remember that these partitions are those corresponding to $m = \tilde{m}$, the observed pattern of missing data. For further notational convenience, we let $\tilde{u} = (u_{(0)}, \tilde{u}_{(1)})$; $\tilde{u}$ consists of a vector of arguments, $u_{(0)}$, corresponding to unobserved random variables, and a vector of known numbers, $\tilde{u}_{(1)} = \tilde{v}_{(1)}$, corresponding to values of observed random variables.

The objective is to use $\tilde{v}$, or equivalently $\tilde{m}$ and $\tilde{u}_{(1)}$, to make inferences about θ. It is common practice to ignore the process that causes missing data when making these inferences. Ignoring the process that causes missing data means proceeding by: (a) fixing the random variable M at the observed pattern of missing data, $\tilde{m}$, and (b) assuming that the values of the observed data, $\tilde{u}_{(1)}$, arose from the marginal density of the random variable $U_{(1)}$:

$$\int f_\theta(u)\, du_{(0)}. \tag{5·1}$$

The central question here concerns the weakest simple conditions on g_ϕ such that ignoring the process that causes missing data will always yield proper inferences about θ.

Three conditions are relevant to answering this question. These conditions place no restrictions on $g_\phi(m|u)$ for values of M other than $\tilde{m}$.

Definition 1. The missing data are missing at random if for each value of ϕ, $g_\phi(\tilde{m}|\tilde{u})$ takes the same value for all $u_{(0)}$.

Definition 2. The observed data are observed at random if for each value of ϕ and $u_{(0)}$, $g_\phi(\tilde{m}|u)$ takes the same value for all $u_{(1)}$.

Definition 3. The parameter ϕ is distinct from θ if their joint parameter space factorizes into a ϕ-space and a θ-space, and when prior distributions are specified for ϕ and θ, if these are independent.

Table 1 classifies the four examples of § 4 in terms of these definitions.

6. Missing data and sampling distribution inference

A sampling distribution inference is an inference that results solely from comparing the observed value of a statistic, e.g. an estimator, test criterion or confidence interval, with the sampling distribution of that statistic under various hypothesized underlying distributions. Within the context of sampling distribution inference, the parameters θ and ϕ have fixed hypothesized values.

Ignoring the process that causes missing data when making a sampling distribution inference about the true value of θ means comparing the observed value of some vector statistic $S(\tilde{v})$, equivalently $S(\tilde{m}, \tilde{u}_{(1)})$, to the distribution of $S(v)$ found from f_θ. More precisely, the sampling distribution of $S(\tilde{v})$ ignoring the process that causes missing data is found by fixing M at the observed $\tilde{m}$ and assuming that the sampling distribution of the observed data follows from density (5·1). The problem with this approach is that for the fixed $\tilde{m}$, the sampling distribution of the observed data, $\tilde{u}_{(1)}$, does not follow from (5·1) which is the marginal density of $U_{(1)}$ but from the conditional density of $U_{(1)}$ given that the random variable M took the value $\tilde{m}$:

$$\int \{f_\theta(u)\, g_\phi(\tilde{m}|u)/k_{\theta,\phi}(\tilde{m})\}\, du_{(0)}, \tag{6·1}$$

where $k_{\theta,\phi}(\tilde{m}) = \int f_\theta(u)\, g_\phi(\tilde{m}|u)\, du$, which is the marginal probability that M takes the value $\tilde{m}$. Hence, the correct sampling distribution of $S(\tilde{v})$ depends in general not only on the fixed hypothesized f_θ but also on the fixed hypothesized g_ϕ.

Theorem 6·1. *Suppose that* (*a*) *the missing data are missing at random and* (*b*) *the observed data are observed at random. Then the sampling distribution of* $S(\tilde{v})$ *under* f_θ *ignoring the process that causes missing data, i.e. calculated from density* (5·1), *equals the correct conditional sampling distribution of* $S(\tilde{v})$ *given* $\tilde{m}$ *under* $f_\theta g_\phi$, *that is calculated from density* (6·1) *assuming* $k_{\theta,\phi}(\tilde{m}) > 0$.

Proof. Under conditions (*a*) and (*b*), for each value of ϕ, $g_\phi(\tilde{m}|u)$ takes the same value for all u; notice that this does not imply U and M are independently distributed unless it holds for all possible $\tilde{m}$. Hence $k_{\theta,\phi}(\tilde{m}) = g_\phi(\tilde{m}|u)$, and thus the distribution of every statistic under density (5·1) is the same as under density (6·1).

Theorem 6·2. *The sampling distribution of* $S(\tilde{v})$ *under* f_θ *calculated by ignoring the process that causes missing data equals the correct conditional sampling distribution of* $S(\tilde{v})$ *given* $\tilde{m}$ *under* $f_\theta g_\phi$ *for every* $S(\tilde{v})$, *if and only if*

$$E_{u_{(0)}}\{g_\phi(\tilde{m}|u)\,|\,\tilde{m}, \tilde{u}_{(1)}, \theta, \phi\} = k_{\theta,\phi}(\tilde{m}) > 0. \tag{6·2}$$

Proof. The sampling distribution of every $S(\tilde{v})$ found from density (5·1) will be identical to that found from density (6·1) if and only if these two densities are equal. This equality may be written as equation (6·2) by dividing by (5·1), and multiplying by $k_{\theta,\phi}(m)$.

The phrase 'ignoring the process that causes missing data when making sampling distribution inferences' may suggest not only calculating sampling distributions with respect to density (6·1) but also interpreting the resulting sampling distributions as unconditional rather than conditional on $\tilde{m}$.

THEOREM 6·3. *The sampling distribution of $S(\tilde v)$ under f_θ calculated ignoring the process that causes missing data equals the correct unconditional sampling distribution of $S(\tilde v)$ under $f_\theta g_\phi$ for all $S(\tilde v)$ if and only if $g_\phi(\tilde m|u) = 1$.*

Proof. The sufficiency is immediate. To establish the necessity consider the statistic $S(v) = 1$ if $m = \tilde m$ and 0 otherwise.

7. MISSING DATA AND DIRECT-LIKELIHOOD INFERENCE

A direct-likelihood inference is an inference that results solely from ratios of the likelihood function for various values of the parameter (Edwards, 1972). Within the context of direct-likelihood inference, θ and ϕ take values in a joint parameter space $\Omega_{\theta,\phi}$.

Ignoring the process that causes missing data when making a direct-likelihood inference for θ means defining a parameter space for θ, Ω_θ, and taking ratios, for various $\theta \in \Omega_\theta$, of the 'marginal' likelihood function based on density (5·1):

$$\mathscr{L}(\theta|\tilde v) = \delta(\theta, \Omega_\theta)\int f_\theta(\tilde u)\, du_{(0)}, \qquad (7\cdot 1)$$

where $\delta(a, \Omega)$ is the indicator function of Ω. Likelihood (7·1) is regarded as a function of θ given the observed $\tilde m$ and $\tilde u_{(1)}$.

The problem with this approach is that M is a random variable whose value is also observed, so that the actual likelihood is the joint likelihood of the observed data $\tilde u_{(1)}$ and $\tilde m$:

$$\mathscr{L}(\theta,\phi|\tilde v) = \delta\{(\theta,\phi), \Omega_{\theta,\phi}\}\int f_\theta(\tilde u)\, g_\phi(\tilde m|\tilde u)\, du_{(0)} \qquad (7\cdot 2)$$

regarded as a function of θ, ϕ given the observed $\tilde u_{(1)}$ and $\tilde m$.

THEOREM 7·1. *Suppose (a) that the missing data are missing at random, and (b) that ϕ is distinct from θ. Then the likelihood ratio ignoring the process that causes missing data, that is $\mathscr{L}(\theta_1|\tilde v)/\mathscr{L}(\theta_2|\tilde v)$, equals the correct likelihood ratio, that is $\mathscr{L}(\theta_1, \phi|\tilde v)/\mathscr{L}(\theta_2, \phi|\tilde v)$, for all $\phi \in \Omega_\phi$ such that $g_\phi(\tilde m|\tilde u) > 0$.*

Proof. Conditions (*a*) and (*b*) imply from equations (7·1) and (7·2) that

$$\mathscr{L}(\theta,\phi|\tilde v) = \delta(\phi, \Omega_\phi)\, g_\phi(\tilde m|\tilde u)\, \mathscr{L}(\theta|\tilde v).$$

THEOREM 7·2. *Suppose $\mathscr{L}(\theta|\tilde v) > 0$ for all $\theta \in \Omega_\theta$. All likelihood ratios for $\theta \in \Omega_\theta$ ignoring the process that causes missing data are correct for all $\phi \in \Omega_\phi$, if and only if (a) $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$, and (b) for each $\phi \in \Omega_\phi$, $E_{u_{(0)}}\{g_\phi(\tilde m|\tilde u)\,|\tilde m, \tilde u_{(1)}, \theta, \phi\}$ takes the same positive value for all $\theta \in \Omega_\theta$.*

Proof. First we show that

$$\mathscr{L}(\theta,\phi|\tilde v) = E_{u_{(0)}}\{g_\phi(\tilde m|\tilde u)\,|\tilde m, \tilde u_{(1)}, \theta, \phi\}\, \delta\{(\theta,\phi), \Omega_{\theta,\phi}\}\, \mathscr{L}(\theta, \tilde v). \qquad (7\cdot 3)$$

This is immediate if $\mathscr{L}(\theta|\tilde v) > 0$ for all $\theta \in \Omega_\theta$, and is true otherwise because

$$\mathscr{L}(\theta|\tilde v) \geqslant \mathscr{L}(\theta,\phi|\tilde v) \geqslant 0$$

for all θ, ϕ and $\tilde v$. If conditions (*a*) and (*b*) hold, (7·2) factorizes into a θ-factor and a ϕ-factor; thus these conditions are sufficient even if $\mathscr{L}(\theta|\tilde v) = 0$ for some $\theta \in \Omega_\theta$.

Now consider the necessity of conditions (*a*) and (*b*). Since $\mathscr{L}(\theta|\tilde v) > 0$ for all $\theta \in \Omega_\theta$, if the likelihood ratios for θ ignoring the process that causes missing data are correct for all $\phi \in \Omega_\phi$, for each $(\theta, \phi) \in \Omega_\theta \times \Omega_\phi$, we have $\mathscr{L}(\theta, \phi | \tilde{v}) > 0$. Hence condition (*a*) in the theorem is necessary. Now using condition (*a*) and (7·3) write for all θ_1, $\theta_2 \in \Omega_\theta$ and $\phi \in \Omega_\phi$

$$\frac{\mathscr{L}(\theta_1, \phi | \tilde{v})}{\mathscr{L}(\theta_2, \phi | \tilde{v})} = \frac{E_{u_{(0)}}\{g_\phi(\tilde{m} | \tilde{u}) \,|\, \tilde{m}, \tilde{u}_{(1)}, \theta_1, \phi\} \,\mathscr{L}(\theta_1 | \tilde{v})}{E_{u_{(0)}}\{g_\phi(\tilde{m} | \tilde{u}) \,|\, \tilde{m}, \tilde{u}_{(1)}, \theta_2, \phi\} \,\mathscr{L}(\theta_2 | \tilde{v})} > 0. \tag{7·4}$$

If (7·4) equals $\mathscr{L}(\theta_1 | \tilde{v}) / \mathscr{L}(\theta_2 | \tilde{v})$ for all $\theta_1, \theta_2 \in \Omega_\theta$ and all $\phi \in \Omega_\phi$, we have condition (*b*) in the theorem.

8. Missing data and Bayesian inference

A Bayesian inference is an inference that results solely from posterior distributions corresponding to specified prior distributions, e.g. the posterior mean and variance of a parameter having a specified prior distribution. Within the context of Bayesian inference, θ and ϕ are random variables whose marginal distribution is specified by the product of the prior densities, $p(\theta)\, p(\phi | \theta)$.

Bayesian inference for θ ignoring the process that causes missing data means choosing $p(\theta)$ and assuming that the observed data, $\tilde{u}_{(1)}$, arose from density (5·1). Hence the posterior distribution of θ ignoring the process that causes missing data is proportional to

$$p(\theta) \int f_\theta(\tilde{u})\, du_{(0)}. \tag{8·1}$$

The problem with this approach is that the random variable M is being fixed at $\tilde{m}$ and thus is being implicitly conditioned upon without being explicitly conditioned upon. That is, correct conditioning on both the observed data, $\tilde{u}_{(1)}$, and on the observed pattern of missing data, $\tilde{m}$, leads to the joint posterior distribution of θ and ϕ which is proportional to

$$p(\theta)\, p(\phi | \theta) \int f_\theta(\tilde{u})\, g_\phi(\tilde{m} | \tilde{u})\, du_{(0)}. \tag{8·2}$$

Theorem 8·1. *Suppose* (*a*) *that the missing data are missing at random, and* (*b*) *that ϕ is distinct from θ. Then the posterior distribution of θ ignoring the process that causes missing data, i.e. calculated from equation* (8·1), *equals the correct posterior distribution of θ, that is calculated from* (8·2), *and the posterior distributions for θ and ϕ are independent.*

Proof. By conditions (*a*) and (*b*), equation (8·2) equals $\{p(\theta) \int f_\theta(\tilde{u})\, du_{(0)}\} \{p(\phi)\, g_\phi(\tilde{m} | \tilde{u})\}$.

Theorem 8·2. *The posterior distribution of θ ignoring the process that causes missing data equals the correct posterior distribution of θ if and only if*

$$E_{\phi, u_{(0)}}\{g_\phi(\tilde{m} | \tilde{u}) | \tilde{m}, \tilde{u}_{(1)}, \theta\} \tag{8·3}$$

takes a constant positive value.

Proof. The posterior distribution of θ is proportional to (8·2) integrated over ϕ. This can be written as

$$\{p(\theta) \int f_\theta(\tilde{u})\, du_{(0)}\} \int E_{u_{(0)}}\{g_\phi(\tilde{m} | \tilde{u}) | \tilde{m}, \tilde{u}_{(1)}, \theta, \phi\}\, p(\phi | \theta)\, d\phi. \tag{8·4}$$

Expressions (8·4) and (8·1) yield the same distribution for θ if and only if they are equal. Hence, the second factor in (8·4), which is expression (8·3), must take a constant positive value.

9. Comparing inferences in a simple example

Suppose that we want to estimate the weight of an object, say θ, using a scale that has a digital display, including a sign bit! The weighing mechanism has a known normal error distribution with mean zero and variance one. We propose to weigh the object ten times and so obtain ten independent, identically distributed observations from $N(\theta, 1)$. A colleague tells us that in his experience sometimes no value will be displayed. Nevertheless in our ten weighings we obtain ten values whose average is 5·0.

Let us first ignore the process that causes missing data. This might seem especially reasonable since there are in fact no missing data. Under f_θ, the sampling distribution of the sample average, 5·0, is $N(\theta, 0{\cdot}1)$, and with a flat prior on $\theta > 0$ the posterior distribution of θ is approximately $N(5{\cdot}0,\ 0{\cdot}1)$. Also, 5·0 is the maximum likelihood estimate of θ, and for example the likelihood ratio of $\theta = 5{\cdot}0$ to $\theta = 4{\cdot}0$ is e^5.

Now let us consider the process that causes missing data. Since there are no missing observations, the missing data are missing at random. We discuss two processes that cause missing data. First suppose that the manufacturer informs us that the display mechanism has the flaw that for each weighing the value is displayed with probability $\phi = \theta/(1+\theta)$. This fact means that the observed data are observed at random, and that ϕ is not distinct from θ. With a flat prior on $\theta > 0$ the posterior distribution for θ is proportional to the posterior distribution ignoring the process that causes missing data times $\{\theta/(1+\theta)\}^{10}$. Thus, because θ and ϕ are not distinct, the posterior distribution for θ may be affected by the process that causes missing data; i.e. all ten weighings yielding values suggests that $\theta/(1+\theta)$ is close to unity and hence suggests that θ is large compared to unity. The maximum likelihood estimate of θ is now about 5·04 and the likelihood ratio of $\theta = 5{\cdot}0$ to $\theta = 4{\cdot}0$ is about $1{\cdot}5\sqrt{e}$.

However, since in this case the missing data are missing at random and the observed data are observed at random, the sampling distribution of the sample average ignoring the process that causes missing data equals the conditional sampling distribution of the sample average given that all values are observed. The unconditional sampling distribution of the sample average is the mixture of eleven distributions, the ith being $N(\theta, 1/i)$ with mixing weight $\theta^i 10!/(1+\theta)^{10}\{i!(10-i)!\}$, and the eleventh being the distribution of the 'sample average' if no data are observed, e.g. zero with probability 1, with mixing weight $(1+\theta)^{-10}$.

Now suppose that the manufacturer instead informs us that the display mechanism has the flaw that it fails to display a value if the value that is going to be displayed is less than ϕ. Then the missing data are still missing at random, but the observed data are not observed at random since the values are observed because they are greater than ϕ. Also θ and ϕ are now distinct since ϕ is a property of the machine and θ is a property of the object. It follows that sampling distribution inferences may be affected by the process that causes missing data. Thus, the sampling distribution of the sample average given that all ten values are observed is now the convolution of ten values from the distribution $N(\theta, 0{\cdot}01)$ truncated below ϕ, and the unconditional sampling distribution of the sample average is the mixture of eleven distributions, the jth $(j = 1, \ldots, 10)$ being the convolution of j $N(\theta, 1/j)$'s with mixing weight equal to $[10!/\{j!(10-j)!\}]\,\xi(\phi,\theta)^j\{1-\xi(\phi,\theta)\}^{10-j}$, where $\xi(\phi,\theta)$ equals the area from ϕ to ∞ under the $N(\theta, 1)$ density, and the eleventh being the distribution of the 'sample average' if no data are observed with mixing weight $\{1-\xi(\phi,\theta)\}^{10}$.

However, since the missing data are missing at random and ϕ is distinct from θ, the posterior distribution for θ with each fixed prior is unaffected by the process that causes missing data. Hence, with a flat prior on $\theta > 0$, the posterior distribution for θ remains approximately $N(5{\cdot}0,\ 0{\cdot}1)$. Also, 5·0 remains the maximum likelihood estimate of θ, and $\sqrt{e}$ remains the likelihood ratio of $\theta = 5{\cdot}0$ to $\theta = 4{\cdot}0$.

10. Practical implications

In order to have a practical problem in mind, consider the example in § 1 of the survey of families in 1967 and the follow-up survey in 1970, where a number of families in the 1967 survey could not be located in 1970. Notice that it may be plausible that the missing data are missing at random; that is, families were not located in 1970 basically because of their values on background variables that were recorded in 1967, e.g. low scores on socioeconomic status measures. Also it may be plausible that the parameter of the distribution of the data and the parameter relating 1967 family characteristics to locatability in 1970 are not tied to each other. However, it is more difficult to believe that the missing data are missing at random and that the observed data are observed at random, because these would imply that families were not located in 1970 independently of both the values that were recorded in 1967 and those that would have been recorded in 1970.

This example seems to suggest that if the process that causes missing data is ignored, Bayesian and direct-likelihood inferences will be proper Bayesian, or likelihood, inferences more often than sampling distribution inferences will be proper sampling distribution inferences. Since explicitly considering the process that causes missing data requires a model for the process, it seems simpler to make proper Bayesian and likelihood inferences in many cases.

One might argue, however, that this apparent simplicity of likelihood and Bayesian inference really buries the important issues. Many Bayesians feel that data analysis should proceed with the use of 'objective' or 'noninformative' priors (Box & Tiao, 1973; Jeffreys, 1961), and these objective priors are determined from sampling distributions of statistics, e.g. Fisher information. In addition, likelihood inferences are at times surrounded with references to the sampling distributions of likelihood statistics. Thus practically, when there is the possibility of missing data, some interpretations of Bayesian and likelihood inference face the same restrictions as sampling distribution inference.

The inescapable conclusion seems to be that when dealing with real data, the practising statistician should explicitly consider the process that causes missing data far more often than he does. However, to do so, he needs models for this process and these have not received much attention in the statistical literature.

I would like to thank A. P. Dempster, P. W. Holland, T. W. F. Stroud and a referee for helpful comments on earlier versions of this paper.

References

Afifi, A. A. & Elashoff, R. M. (1966). Missing observations in multivariate statistics. I. Review of the literature. *J. Am. Statist. Assoc.* **61**, 595–604.

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

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, Mass: Addison-Wesley.

Cochran, W. G. (1963). *Sampling Techniques*. New York: Wiley.

Edwards, A. W. F. (1972). *Likelihood*. Cambridge University Press.

Hartley, H. O. (1956). Programming analysis of variance for general purpose computers. *Biometrics* **12**, 110–22.

Hartley, H. O. & Hocking, R. R. (1971). Incomplete data analysis. *Biometrics* **27**, 783–823.

Healy, M. J. R. & Westmacott, M. (1956). Missing values in experiments analyzed on automatic computers. *Appl. Statist.* **5**, 203–6.

Hocking, R. R. & Smith, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. *J. Am. Statist. Assoc.* **63**, 159–73.

HOCKING, R. R. & SMITH, W. B. (1972). Optimum incomplete multi-normal samples. *Technometrics* **14**, 299–307.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd edition. Oxford: Clarendon.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. New York: Wiley.

RUBIN, D. B. (1972). A noniterative algorithm for least squares estimation of missing values in any analysis of variance design. *Appl. Statist.* **21**, 136–41.

RUBIN, D. B. (1975). Bayesian inference for causality: The importance of randomization. *Proc. Social Statistics Section, Am. Statist. Assoc.* pp. 233–9.

RUBIN, D. B. (1976). Noniterative least squares estimates, standard errors, and *F*-tests for analyses of variance with missing data. *J. R. Statist. Soc.* B **38**. To appear.

TRAWINSKI, I. M. & BARGMANN, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. *Ann. Math. Statist.* **35**, 647–57.

WILKINSON, G. N. (1958). Estimation of missing values for the analysis of incomplete data. *Biometrics* **14**, 257–86.

WILKS, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Ann. Math. Statist.* **3**, 163–95.

[*Received April* 1974. *Revised November* 1975]

Comments on paper by D. B. Rubin

BY R. J. A. LITTLE

Department of Statistics, University of Chicago

In the following comments, a notation close to that of Dr Rubin's paper is used. Thus $U = (u_1, \ldots, u_n)$ denotes the full data, with density $f(u;\theta)$ $(\theta \in \Omega_\theta)$ and $M = (m_1, \ldots, m_n)$ indicates the observed pattern, with conditional density $g(m|u;\phi)$ $(\phi \in \Omega_\phi)$ given $U = u$. The distribution of obs(U, M), the observed data, can be described as follows. It has $M = m$ with probability

$$g(m;\, \theta, \phi) = \int g(m|u;\phi) f(u;\theta)\, du = E_U\{g(m|U;\, \phi);\, \theta\}. \tag{1}$$

Given $M = m$, the conditional density of obs(U, M) is

$$f(u_{(1)}|m;\, \theta, \phi) = f(u_{(1)};\, \theta) \int g(m|u;\, \phi) f(u_{(1)}|u_{(0)};\, \theta)\, du_{(0)} \tag{2}$$

$$= f(u_{(1)};\, \theta) \frac{E_{U_{(0)}}\{g(m|U_{(0)}, u_{(1)};\, \phi)\,|u_{(1)};\, \theta\}}{E_U\{g(m|U;\, \phi);\, \theta\}}, \tag{3}$$

where $U_{(1)}$ is the observed part of U and $U_{(0)}$ is the missing part of U.

For sampling based inferences, a first crucial question concerns when it is justified to condition on the observed pattern, that is on the event $M = m$, and to use the distribution (2) and (3). A natural condition is that M should be ancillary, that is that $g(m;\, \theta, \phi)$ should be independent of θ for all m, ϕ. Otherwise the pattern on its own carries at least some information about θ, which should in principle be used.

Suppose now that this ancillarity condition is satisfied. As Dr Rubin stresses, ignoring the deletion mechanism involves not only conditioning on $M = m$, but also assuming that $U_{(1)}$ has a distribution with marginal density $f(u_{(1)};\, \theta)$, that is that for the observed pattern $M = \tilde{m}$,

$$f(u_{(1)}|\tilde{m};\, \theta, \phi) = f(u_{(1)};\, \theta), \tag{4}$$

or that $g(\tilde{m}|u_{(1)};\, \theta, \phi) = E_{U_0}\{g(\tilde{m}|U_{(0)}, u_{(1)};\, \theta\}$ is independent of $u_{(1)}$, which is Dr Rubin's condition (6·2).

A sufficient condition for (4) is a combination of Dr Rubin's conditions, missing at random and observed at random, namely that

$$f(\tilde{m}|u;\, \phi) \text{ is independent of } u, \tag{5}$$

This implies ancillarity if and only if it holds for all observable patterns m, and not just for the observed pattern $\tilde{m}$, and also the parameter space for (θ, ϕ) is $\Omega_\theta \times \Omega_\phi$; then the deletion pattern can be ignored. For example, consider Dr Rubin's weighing problem in § 9, when a weighing value is displayed with probability $\theta/(1-\theta)$, and all values are displayed. Then (5) is satisfied for all patterns m, but $\theta = \phi$, so that θ and ϕ are dependent, and ancillarity fails to hold. Thus in principle the rather complicated distribution of obs(U, M) described by Dr Rubin should be used. However this deletion mechanism seems highly unlikely in practice.

Necessary conditions for ignoring the deletion mechanism are unfortunately not obvious, and it is worth considering some further examples.

Example 1. Suppose that for the observed value $\tilde{m}$, $U_{(0)}$ and $U_{(1)}$ are independently distributed, and that the probability that $M = \tilde{m}$ depends on $U_{(0)}$ but not $U_{(1)}$, that is $g(\tilde{m}|u; \phi) = g(\tilde{m}|u_{(0)}; \phi)$. Then clearly (4) is satisfied but not (5), so (5) is not necessary for (4).

Example 2. Let U_i be independent $N(\theta, 1)$ $(i = 1, \ldots, n)$ and suppose $m_i = 1$ if and only if $|U_i - \bar{U}| < \phi$, for some constant ϕ. A simple computation of (1) establishes that m is ancillary for θ. However we cannot ignore the deletion mechanism, since the correct distribution for sampling inference has density

$$f(u_{(1)}|\tilde{m}; \theta, \phi) = \frac{\int_{R(\tilde{m})} f(u_{(0)}|u_{(1)}; \theta)\, du_{(0)}}{\int_{R(\tilde{m})} f(u; \theta)\, du}, \qquad (3)$$

where $R(\tilde{m}) = \{u\colon |u_i - \bar{u}| \gtrless k \text{ as } \tilde{m}_i = 0 \text{ or } 1\}$ is a region of R^n; this is clearly not the normal density $f(u_{(1)}; \theta)$.

The case of pure likelihood inferences is much simpler, since we can fix $U_{(1)}$ and M at their observed values $\tilde{u}_1, \tilde{m}$, and the rather complex sample space of obs (U, M) is not relevant. Dr Rubin's sufficient conditions in Theorem 7·1 are perhaps more remarkable than his examples would suggest. His Example 3 for instance, is already well known: see Examples 2·34 and 2·40 of Cox & Hinkley (1974). We give a multivariate example of some practical importance.

Example 3. Consider an incomplete bivariate normal sample size n of random variables X and Y, which have respective means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation ρ. Suppose X is always observed. Two possible deletion mechanisms for Y are: (a) observe Y if and only if $Y > c$; (b) observe Y if and only if $X > c$. It is easily seen that Dr Rubin's 'missing at random' condition is satisfied in (b) but not in (a), and so for maximum likelihood estimation we can ignore the deletion mechanism in (b) but not in (a). To illustrate this, the estimates of Table 1 were found from generated data with 50 observations, $c = 0$ and $\mu_1 = \mu_2 = 0$, so that about half the Y values were deleted in (a) and (b). Note that estimates of μ_2, σ_2^2 and ρ in situation (ii$a$) are biased, confirming previous theory. However the estimates in situation (ii$b$) are maximum likelihood, and are close to their true values. Thus here we can ignore the deletion pattern, although the observed values of Y do not follow the marginal $N(0, 2)$ distribution, and in particular their sample mean will overestimate zero.

In a real set of data for which (ii$b$) is appropriate, X might be blood pressure, and Y a medical test which for safety reasons is not carried out when X is below a certain level c.

Table 1. *Maximum likelihood estimates, ignoring the deletion mechanism, for* $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 2$, $\rho = 0·71$

		$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\rho}$
(i)	Complete data	0·013	0·085	0·917	1·827	0·780
(ii$a$)	Data censored by (a)	0·013	0·930	0·917	0·456	0·510
(ii$b$)	Data censored by (b)	0·013	−0·140	0·917	1·991	0·645

In summary, Dr Rubin's paper should stimulate thought about the many mechanisms which produce data with missing values.

Reference

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Reply to comments

By D. B. RUBIN

First, I want to thank Dr Little for his Example 3, which numerically illustrates the point being made in the beginning of § 10. Secondly, I must reject his restriction that M should be ancillary when making sampling distribution inferences for θ which are conditional on M. As Theorem 6·1 states, if (a) the missing data are missing at random and (b) the observed data are observed at random, then a sampling distribution probability statement that ignores the process that causes missing data is correct if interpreted as being conditional on M. Given (*a*) and (*b*), Theorem 7·1 on likelihood inference implies that such a probability statement cannot generally be fully efficient for inference about θ unless (*c*) θ is distinct from ϕ. Nevertheless, sampling distribution inferences that are less than fully efficient are often quite useful. Furthermore, given (*a*), (*b*) and (*c*), sampling distribution inference for θ should be conditional on M whether or not M is ancillary. For a simple case, consider my Example 4 with $\tilde{m} = (1, 0)$, $\phi = 0{\cdot}1$, and $(u_1, u_2) \sim N\{(\theta, \theta), I\}$. The conditional probability of the event $\mathscr{E} = (\bar{u} - 1{\cdot}96 < \theta < \bar{u} + 1{\cdot}96)$, where $\bar{u} = \Sigma m_i u_i / \Sigma m_i$, is 0·95 for all θ, while the unconditional probability of $\mathscr{E}$ is nearly 0·99 for θ quite positive. This example suggests that the usual definition of ancillary (Cox & Hinkley, 1974, p. 35) is incorrect for inference about θ and should be modified to be conditional on the observed value of the ancillary statistic.