RB-75-14

RESEARCH BULLETIN

INFERENCE AND MISSING DATA

Donald B. Rubin

This paper has been accepted for publication by Biometrika. Any citations should be made to the Journal rather than to the Bulletin.

Educational Testing Service
Princeton, New Jersey
April 1975

INFERENCE AND MISSING DATA

Donald B. Rubin
Educational Testing Service

Abstract

Two results are presented concerning inference when data may be missing. First, ignoring the process that causes missing data when making sampling distribution inferences about the parameter of the data, θ, is generally appropriate if and only if the missing data are "missing at random" and the observed data are "observed at random," and then such inferences are generally conditional on the observed pattern of missing data. Second, ignoring the process that causes missing data when making Bayesian inferences about θ is generally appropriate if and only if the missing data are missing at random and the parameter of the missing data is "independent" of θ. Examples and discussion indicating the implications of these results are included.

Some key words: missing data, missing values, incomplete data, observed at random, missing at random, sampling distribution inference, likelihood inference, Bayesian 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 and Elashoff, 1966; Anderson, 1957; Hartley and Hocking, 1971; Hocking and 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 and Westmacott, 1956; Rubin, 1972; 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 and Smith, 1972; Trawinski and Bargmann, (1964), random sampling from a finite population (the values of variables for unsampled units being missing, Cochran, 1963, p. 18), randomization in an experiment (for each unit, the values that would have been observed had the unit received a different treatment are missing, Kempthorne, 1952, p. 137), sequential stopping rules (the values after the last one observed are missing, Lehmann, 1959) and even some "robust analyses" (observed values are considered outliers and so discarded or made missing).

Our view of missing data is very encompassing and can include most problems of statistical inference. That is, one can consider the desired data set to be that data set from which one could simply calculate the desired summaries. One rarely observes this ideal data set but a reduced one with missing data. The problem of statistical inference can thus be thought of as the problem of inference about the distribution of missing data.

2. Objective and Overview

Our objective is to find the weakest 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 presents 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 dis-

tribution of the missing-data indicator given the data). Section 4 presents examples of processes that cause missing data. Section 5 defines ignoring the process that causes missing data to correspond to ignoring the fact that the missing data indicator is itself a random variable whose value is observed.

Section 6 defines the three relevant conditions on the process that causes missing data and relates them to the examples presented in Section 3. The first two are conditional on the observed value of the missing-data indicator. The definitions correspond to the following intuitive statements.

> The missing data are missing at random if, given the values of the observed data and the parameter ϕ, the values of the missing data did not influence the observed pattern of missing data.
>
> The observed data are observed at random if, given the values of the missing data and the parameter ϕ, the values of the observed data did not influence the observed pattern of missing data.
>
> The parameter ϕ is independent of θ if the a priori restrictions on ϕ are the same for each possible value of θ.

Section 7 states and proves Theorems 1 and 2 which show that ignoring the process that causes missing data when making sampling distribution inferences about θ is generally appropriate if and only if the missing data are missing at random and the observed data are observed at random, and then such inferences are generally conditional on the observed pattern of missing data.

Section 8 states and proves Theorem 3 which shows that ignoring the process that causes missing data when making Bayesian (or likelihood) inferences about θ is generally appropriate if and only if the missing data are missing at random and ϕ is independent of θ.

The reader not interested in the formal details should be able to skim Sections 3-8 and proceed to Section 9.

Section 9 uses these results to highlight the distinction between the sampling distribution and 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.

3. Notation for the Random Variables

In order to be unambiguous, the notation here has to be precise. We will use boldface to indicate a random variable (normal face indicating the values of the random variable), upper case to indicate vectors (lower case indicating scalars), and a tilda (~) to indicate a specific observed value.

Let $\underline{\mathbf{U}} = (\underline{\mathbf{u}}_1, \ldots, \underline{\mathbf{u}}_n)$ be a vector random variable (r.v.) with probability density function (p.d.f.) f_θ. The data analyst's objective is to make inferences about θ, the vector parameter of this density. Often in practice, the r.v. $\underline{\mathbf{U}}$ will be arranged in a "units" by "variables" matrix. Let $\underline{\mathbf{M}} = (\underline{\mathbf{m}}_1, \ldots, \underline{\mathbf{m}}_n)$ be the associated "missing-data indicator" vector r.v. The probability that $\underline{\mathbf{M}}$ takes the value $M = (m_1, \ldots, m_n)$ given that $\underline{\mathbf{U}}$ takes the value $U = (u_1, \ldots, u_n)$ is $g_\phi(M;U)$ where ϕ is the nuisance vector parameter of the p.d.f.

The conditional p.d.f. g_ϕ corresponds to "the process that causes missing data": if $m_i=1$, the value of the r.v. $\underline{\mathbf{u}}_i$ will be observed while if $m_i=0$, the value of $\underline{\mathbf{u}}_i$ will not be observed. More precisely, define the vector r.v. $\underline{\mathbf{V}} = (\underline{\mathbf{v}}_1, \ldots, \underline{\mathbf{v}}_n)$ with range extended to include the special value $\square$ for missing data:

$$v_i = \begin{cases} u_i & \text{if } m_i=1 \\ \square & \text{if } m_i=0 \end{cases} .$$

The data analyst observes the values of the r.v. $\underline{V}$, not the r.v. $\underline{U}$, although he wishes to make inferences about the distribution of $\underline{U}$.

4. Examples of Processes That Cause Missing Data

In order to clarify the notation in Section 3 we present four examples.

Example 1: Equally likely missing values. 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: Censored data. Let u_i be the value of blood pressure for the ith subject, i=1,...,n in a hospital survey. Suppose $v_i = \square$ 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 $a \geq 0$, 0 otherwise

$\delta(a) = 1$ if $a=0$, 0 otherwise.

Example 3: Sequential sampling. Observations are taken in sequence until a specified function of the observed data is in a specified critical region (n is essentially infinite and for some n_0 which is a function of the observed data $v_i \neq \square$ if $i \leq n_0$ and $v_i = \square$ if $i > n_0$). Thus

$$g_\phi(M;U) = \left[\prod_{i=1}^{n_0} \delta(1-m_i)\right]\left[\prod_{i=n_0+1}^{n} \delta(m_i)\right]$$

where n_0 = minimum k such that the function $Q_k(u_1,\ldots,u_k) \in C$.

Example 4: A small but complex case. Let n=2. If $u_1 \geq 0$: with probability ϕ $v_1 \neq \square$ and $v_2 = \square$, and with probability $1-\phi$ $v_1 \neq \square$ and $v_2 \neq \square$. If $u_1 < 0$: with probability ϕ $v_1 \neq \square$ and $v_2 = \square$, and with probability $1-\phi$ $v_1 = \square$ and $v_2 \neq \square$. 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 $\tilde{V} = (\tilde{v}_1,\ldots,\tilde{v}_u)$ be a particular sample realization of $\underline{V}$, i.e., each $\tilde{v}_i$ is either a known number or a missing value, $\square$. These observed values imply an observed value for the r.v. $\underline{M}$, $\tilde{m} = (\tilde{m}_1,\ldots,\tilde{m}_n)$, and imply observed values for some of the scalar r.v.'s in $\underline{U}$. That is, if $\tilde{v}_i$ is a number, then the observed value of $\underline{m}_i$, $\tilde{m}_i$, is 1 and the observed value of $\underline{u}_i$, $\tilde{u}_i$, is $\tilde{v}_i$; if $\tilde{v}_i = \square$, then $\tilde{m}_i = 0$ and the value of $\underline{u}_i$ is not known (in special cases, knowing values in $\tilde{V}$ may imply values for some $\underline{u}_i$ with $\tilde{v}_i = \square$; e.g., f_θ specifies $u_1 = u_2 + u_3$ and we observe $\tilde{v}_1 = \square$, $\tilde{v}_2 = 3.1$, $\tilde{v}_3 = 5.2$).

The data analyst's objective is to use $\tilde{V}$ (equivalently, $\tilde{M}$ and $\{\tilde{u}_i | \tilde{m}_i = 1\}$), 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 r.v. $\underline{M}$ at the observed pattern of missing data, $\tilde{M}$, and (b) assuming that the observed data $\{\tilde{v}_i | \tilde{m}_i = 1\} = \{\tilde{u}_i | \tilde{m}_i = 1\}$ arose from the marginal p.d.f. of the r.v.'s $\{\underline{u}_1 | \tilde{m}_i = 1\}$:

$$(1) \qquad \int f_\theta(U)\, dU\{\tilde{m}_i = 0\}$$

where $$dU\{\tilde{m}_i = 0\} = \prod_{\{i | \tilde{m}_i = 0\}} du_i \quad .$$

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

6. Three Definitions

Three simple conditions are relevant to answering this question. The first two conditions are on $g_\phi(\tilde{M};U)$ which is the p.d.f. of $\underline{M}$ given $\underline{U}$ evaluated at the observed fixed value of $\underline{M}$, $\tilde{M}$, regarded as a function of ϕ and U. Hence, these two conditions place no restrictions on $g_\phi(M,U)$ for values of $\underline{M}$ other than $\tilde{M}$. Table 1 classifies the four examples of Section 3 in terms of these definitions.

In these definitions and in Sections 7 and 8 we will use the expression "functionally independent." Some function of X and Y Z(X,Y) is functionally independent of X if $Z(X,Y) = Z^*(Y)$ for all values of Y and some function Z^*.

Definition 1: The missing data are missing at random if $g_\phi(\tilde{M};U)$ is functionally independent of all $\tilde{u}_i$ such that $\tilde{m}_i = 0$.

Definition 2: The observed data are observed at random if $g_\phi(\tilde{M};U)$ is functionally independent of all $\tilde{u}_i$ such that $\tilde{m}_i=1$.

These two conditions are relevant for sampling distribution inferences about θ as discussed in Section 6. Notice that if the missing data are missing at random and the observed data are observed at random, then $g_\phi(\tilde{M};U)$ is functionally independent of U, and if this holds for all possible $\tilde{M}$, then the r.v.'s $\underline{M}$ and U are independently distributed.

Definitions 1 and 3 are relevant for Bayesian inferences as discussed in Section 8.

Definition 3: The parameter ϕ is independent of θ if the a priori restrictions on the values of ϕ are the same for every possible θ, i.e., the prior distributions of ϕ and θ are independent.

If the missing data are missing at random and ϕ is independent of θ, then $g_\phi(\tilde{M};U)$ is independent of the "relevant unknowns": θ and the missing data.

TABLE 1: Classifying the Examples in Section 3

Example	Missing Data Missing at Random	Observed Data Observed at Random	ϕ Independent of θ
1	Always MAR	Always OAR	Always independent
2	MAR only if all $\tilde{m}_i=1$	OAR only if all $\tilde{m}_i=0$	Never independent
3	Always MAR	Never OAR	Always independent
4	MAR unless $\tilde{M}=(0,1)$	OAR unless $\tilde{M}=(1,1)$	Independent if a priori ϕ is not restricted by θ.

7. Missing Data and Sampling Distribution Inferences About θ

We now show that ignoring the process that causes missing data when making sampling distribution inferences about θ is generally appropriate if and only if the missing data are missing at random and the observed data are observed at random. In this case, these inferences are conditional on the observed pattern of missing data and are not in general the same as unconditional sampling distribution inferences.

By a sampling distribution inference about θ, we mean an inference about θ that can be interpreted as resulting solely from sampling distributions of statistics; for example, the sampling distribution of an estimator of θ, a test statistic for θ, or a confidence interval for θ. Other examples of sampling distribution inferences include inferences that are based on theorems about the admissibility and efficiency of estimators, as well as the power and levels of tests.

Consider some statistic $S(\tilde{V})$ computed from $\tilde{V}$ (equivalently, some statistic computed from $\tilde{M}$ and the observed data $\{\tilde{v}_i | \tilde{m}_i=1\} = \{\tilde{u}_i | \tilde{m}_i=1\}$). The sampling distribution of $S(\tilde{V})$ ignoring the process that causes missing data is found by fixing $\underline{M}$ at the observed $\tilde{M}$ and assuming that the sampling distribution of the observed data is given by p.d.f. (1). The problem with this approach is that for the fixed $\tilde{M}$, the sampling distribution of the observed data, $\{\tilde{u}_i | \tilde{m}_i = 1\}$, is not p.d.f. (1) which is the marginal p.d.f. of $\{\underline{u}_i | \tilde{m}_i=1\}$ but the conditional p.d.f. of $\{\underline{u}_i | \tilde{m}_i= 1\}$ given that the r.v. $\underline{M}$ took the value $\tilde{M}$:

$$(2) \qquad \int h_{\theta,\phi}(U;\tilde{M}) \; dU\{\tilde{m}_i = 0\}$$

where

$$h_{\theta,\phi}(U;M) = f_\theta(U)\ g_\phi(M;U)/k_{\theta,\phi}(M)$$

and $$k_{\theta,\phi}(M) = \int f_\theta(U)\ g_\phi(M;U) \prod_{i=1}^{n} du_i$$

= the marginal probability that $\underline{M}$ takes the value M.

The following result gives the weakest conditions on g_ϕ such that for every distribution f_θ and every statistic $S(\tilde{V})$ it is appropriate to ignore the process that causes missing data when making sampling distribution inferences about θ.

Theorem 1: The sampling distribution of the statistic $S(\tilde{V})$ calculated by ignoring the process that causes missing data equals the conditional sampling distribution of $S(\tilde{V})$ given $\tilde{M}$ for every f_θ and every statistic $S(\tilde{V})$ if and only if

(a) the missing data are missing at random, and

(b) the observed data are observed at random.

Proof:

The sampling distribution of every statistic $S(\tilde{V})$ found from p.d.f. (1) will be identical to that found from p.d.f. (2) if and only if these two p.d.f.'s are identical. This equality may be rewritten as

$$(3) \qquad E_\theta^{\tilde{M}}[g_\phi(\tilde{M};U)] = k_{\theta,\phi}(\tilde{M})$$

where $$E_\theta^{\tilde{M}}[g_\phi(\tilde{M};U)] = \int g_\phi(\tilde{M};U)\ f_\theta(U)\ dU\{\tilde{m}_i=0\} / \int f_\theta(U)dU\{\tilde{m}_i=0\}$$

is the conditional expectation of $g_\phi(\tilde{M};U)$ over the p.d.f. of the r.v.'s $\{\underline{u}_i|\tilde{m}_i=0\}$ given the r.v.'s $\{\underline{u}_i|\tilde{m}_i=1\}$. Hence, for p.d.f. (1) to equal p.d.f. (2), $E_\theta^{\tilde{M}}[g_\phi(\tilde{M};U)]$ must be functionally independent of U. This will occur for all f_θ if and only if $g_\phi(\tilde{M};U)$ is functionally independent of U.

The phrase "ignoring the process that causes missing data when making sampling distribution inferences" may suggest not only calculating sampling distributions with respect to p.d.f. (1) but also interpreting the resulting sampling distributions as unconditional rather than conditional on $\tilde{M}$. The unconditional sampling distribution of $S(\tilde{V})$ is the distribution of $S(V)$ where $\{\underline{u}_i | m_i=1\}$ and $\underline{M}$ have joint p.d.f.

$$(4) \qquad \int f_\theta(U) g_\phi(M;U) dU \{m_i=0\} = k_{\theta,\phi}(M) \int h_{\theta,\phi}(U;M)\, dU\{m_i=0\} \quad .$$

The following theorem holds for every f_θ.

Theorem 2: The sampling distribution of the statistic $S(\tilde{V})$ calculated by ignoring the process that causes missing data equals the unconditional sampling distribution of $S(\tilde{V})$ for every $S(\tilde{V})$ if and only if $g_\phi(\tilde{M};U)=1$; i.e., $\tilde{M}$ is the only possible pattern of missing data.

Proof: The sufficiency of the condition $g_\phi(\tilde{M};U)=1$ for every f_θ and $S(\tilde{V})$ is obvious. For each f_ϕ, its necessity for all statistics follows immediately from considering the statistic which is 1 if $M = \tilde{M}$ and is zero otherwise.

8. Missing Data and Bayesian Inferences About θ

We now show that ignoring the process that causes missing data when making Bayesian inferences about θ is generally appropriate if and only if the missing data are missing at random and ϕ is independent of θ.

By a Bayesian inference about θ we mean an inference about θ that can be interpreted as resulting solely from posterior distributions for θ corresponding to specified priors. Clearly, the posterior mean and variance of θ with a specified prior on θ are Bayesian inferences. However, this definition of Bayesian inference

is intended to be quite general and includes inferences some might rather call direct likelihood-based (cf Edwards, 1972). For example, the maximum likelihood estimate (m.l.e.) of $\theta\varepsilon\Omega_\theta$ is a Bayesian inference in the above sense since it is the mode of the posterior distribution of θ with prior locally flat on Ω_θ and zero outside Ω_θ. In addition, the likelihood ratio test (l.r.t.) of $\theta\varepsilon\Omega_\theta$ vs. $\theta\varepsilon\Omega_\theta' \subset \Omega_\theta$ is a Bayesian inference since it is the ratio of the maximum of the posterior for θ with prior locally flat on Ω_θ and zero outside and the maximum of the posterior for θ with prior locally flat on Ω_θ' and zero outside.

Consider some prior distribution for θ, say $p(\theta)$. The posterior distribution for θ ignoring the process that causes missing data is proportional to the product of $p(\theta)$ and $\mathcal{L}(\theta;\tilde{V})$, the likelihood function of θ ignoring the process that causes the missing data. This function is the likelihood of the observed data $\{\tilde{u}_i|\tilde{m}_i=1\}$ calculated from p.d.f. (1), i.e.,

$$(5) \qquad \mathcal{L}(\theta;\tilde{V}) = \int f_\theta(U)dU \{\tilde{m}_i=0\}\Big|_{u_i=\tilde{u}_i}$$

regarded as a function of θ for the fixed $\{\tilde{u}_i|\tilde{m}_i=1\}$ and $\tilde{M}$. The problem with this approach from a Bayesian point of view is that the r.v. $\underline{M}$ is being fixed at $\tilde{M}$ and thus is being implicitly conditioned upon without being explicitly conditioned upon. In other words, the likelihood that the r.v. $\underline{V}$ took the value $\tilde{V}$ is the joint likelihood of the observed data $\{\tilde{u}_i|\tilde{m}_i=1\}$ and $\tilde{M}$:

$$(6) \qquad \mathcal{L}(\theta,\phi;\tilde{V}) = \int f_\theta(U)g_\phi(\tilde{M};U)\, dU\{\tilde{m}_i=0\}\Big|_{u_i=\tilde{u}_i}$$

regarded as a function of θ,ϕ for the fixed $\{\tilde{u}_i|\tilde{m}_i=1\}$ and $\tilde{M}$.

Hence, the joint posterior of θ and ϕ is proportional to

$$(7) \qquad [p(\theta)\mathcal{L}(\theta;\tilde{V})]\,[p(\phi;\theta)\mathcal{L}(\theta,\phi;\tilde{V})/\mathcal{L}(\theta;\tilde{V})]$$

where the first factor in (7) is proportional to the posterior of θ ignoring the process that causes missing data.

The following theorem, which holds for every prior $p(\theta)$, gives the weakest conditions on g_ϕ such that for every distribution f_θ it is appropriate to ignore the process that causes missing data when making Bayesian inferences about θ.

Theorem 3: The posterior distribution of θ calculated by ignoring the process that causes missing data equals the posterior distribution of θ for every f_θ if and only if

(a) the missing data are missing at random and

(b) ϕ is independent of θ.

Proof: The posterior of θ ignoring the process that causes missing data will equal the posterior of θ (marginal and conditional given every value of ϕ) if and only if the second factor in (7) is functionally independent of θ. But this factor may be written as

$$(8) \qquad p(\phi;\theta)\, E_\theta^{\tilde{M}}\,[g_\phi(\tilde{M};U)]\Big|_{u_i=\tilde{u}_i}$$

which is functionally independent of θ for all f_θ if and only if $p(\phi;\theta)$ is functionally independent of θ and $g_\phi(\tilde{M};U)$ is functionally independent of $\{u_i|\tilde{m}_i=0\}$.

9. Comparison of Sampling Distribution and Bayesian Inferences

By Theorems 1,2, and 3, if the missing data are not missing at random, generally both sampling distribution and Bayesian inferences are affected by the process that causes missing data. Moreover, even if the missing data are missing at random, inferences can be affected by the process that causes missing data. In order to demonstrate this and the differences between the sampling distribution and Bayesian approaches to the problem of missing data, we consider a very simple example in which the missing data are missing at random but the observed data may or may not be observed at random and ϕ may or may not be independent of θ.

Suppose 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 milligram (mg.). 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 mg.

Let us first ignore the process that causes missing data. This might seem especially reasonable since there are no missing data. Under f_θ, the sampling distribution of the sample average, 5 mg., is $N(\theta,.1)$, and the posterior distribution of θ with a locally flat prior on θ is $N(5,.1)$. Hence, 5 is the m.l.e. of θ and for example the l.r.t. of $\theta=5$ vs. $\theta=4$ is $\sqrt{e}$.

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 ϕ equal to $\frac{\theta}{(1+\theta)}$. This fact means that the observed data are observed at random, but ϕ is not independent of θ. The posterior distribution for θ with locally flat prior on θ is proportional to the posterior ignoring the process that causes missing data times $(\theta/\ 1+\theta)^{10}$ but zero if $\theta \leq 0$. Thus, because θ and ϕ are not independent, the posterior for θ is 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 m.l.e. of θ is now about 5.04 and the l.r.t. of $\theta=5$ vs. $\theta=4$ is about 1.5 $\sqrt{e}$.

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 $\binom{10}{i}\ \theta^i/(1+\theta)^{10}$, 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 missing data are still missing at random, but the observed data are not observed at random since they are observed because they are greater than ϕ. Also θ and ϕ are now independent. It follows that sampling distribution inferences are affected by the process that causes missing data. Thus in our example, the sampling distribution of the sample average given that all ten values are observed is now the convolu-

tion of ten $N(\theta, .01)$'s truncated below ϕ, and the unconditional sampling distribution of the sample average is the mixture of eleven distributions, the ith (i=1,...,10) being the convolution of i $N(\theta, 1/i^2)$'s with mixing weight equal to $\binom{10}{i}\xi(\phi,\theta)^i[1-\xi(\phi,\theta)]^{10-i}$ where $\xi(\phi,\theta)$ = 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 independent of θ, the posterior for θ with any fixed prior is unaffected by the process that causes missing data. Hence, with flat prior on θ, the posterior for θ remains $N(5,.1)$ and 5 remains the m.l.e. of θ and $\sqrt{e}$ remains the l.r.t. of $\theta=5$ vs. $\theta=4$.

10. Practical Implications

In order to have a practical problem in mind, consider the example in Section 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 is 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 is 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 in any way. However, it seems 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 variables that were recorded in 1967 and those that would have been recorded in 1970.

This example seems to imply that if the process that causes missing data is ignored, Bayesian inferences will be proper Bayesian inferences more often than sampling distribution inferences will be proper sampling distribution inferences. Hence, in many practical problems with unintended missing data, if sampling distribution inferences are to be made, the process that causes missing data should be considered explicitly. Since explicitly considering this process requires a model for the process, it seems simpler to make proper Bayesian and likelihood inferences in such cases.

One might argue, however, that this apparent simplicity of Bayesian inference really buries the important issues. Many Bayesians feel that data analysis should proceed with the use of "objective" or "noninformative" priors (Box and Tiao, 1973; Jefferys, 1961). These objective priors are determined from unconditional sampling distributions of statistics (e.g., Fisher information) and thus by Theorem 2, cannot be found in general without explicitly considering the process that causes missing data.

The inescapable conclusion seems to be that when dealing with real data, the practicing 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.

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.

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: John Wiley and Sons.

Edwards, A.W.F. (1972). Likelihood. New York: Cambridge University Press.

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. Biometrics, 27, 783-823.

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.

Jefferys, H. (1961). Theory of Probability, third edition. Oxford: Clarendon Press.

Kempthorne, O. (1952). The Design and Analysis of Experiments. New York: John Wiley & Sons.

Lehmann, E.L.(1959). Testing Statistical Hypotheses. New York: John Wiley & Sons.

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.

Rubin, D.B. (1974). Characterizing the estimation of parameters in incomplete-data problems. Journal of American Statistical Association, 69, 467-474.

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.