What Is Meant by "Missing at Random"?

Author(s): Shaun Seaman, John Galati, Dan Jackson and John Carlin

Source: *Statistical Science*, May 2013, Vol. 28, No. 2 (May 2013), pp. 257-268

Published by: Institute of Mathematical Statistics

Stable URL: https://www.jstor.org/stable/43288491

Statistical Science
2013, Vol. 28, No. 2, 257–268
DOI: 10.1214/13-STS415

What Is Meant by "Missing at Random"?

Shaun Seaman, John Galati, Dan Jackson and John Carlin

Abstract. The concept of missing at random is central in the literature on statistical analysis with missing data. In general, inference using incomplete data should be based not only on observed data values but should also take account of the pattern of missing values. However, it is often said that if data are missing at random, valid inference using likelihood approaches (including Bayesian) can be obtained ignoring the missingness mechanism. Unfortunately, the term "missing at random" has been used inconsistently and not always clearly; there has also been a lack of clarity around the meaning of "valid inference using likelihood". These issues have created potential for confusion about the exact conditions under which the missingness mechanism can be ignored, and perhaps fed confusion around the meaning of "analysis ignoring the missingness mechanism". Here we provide standardised precise definitions of "missing at random" and "missing completely at random", in order to promote unification of the theory. Using these definitions we clarify the conditions that suffice for "valid inference" to be obtained under a variety of inferential paradigms.

Key words and phrases: Ignorability, direct-likelihood inference, frequentist inference, repeated sampling, missing completely at random.

1. INTRODUCTION

The literature on missing data is not entirely clear with respect to the assumptions required for different types of analysis to be valid. First, although the term "missing at random" (MAR) has been widely regarded as central to the theory underlying missing data methods since the seminal paper of Rubin (1976) [33], it has not always been used in a consistent manner. There has often been a lack of detail about whether the MAR condition is a statement only about the realised missingness pattern or about all possible patterns and whether it is only about the realised values of the observed data or all possible observable data values. Second, the distinction between direct-likelihood and frequentist inference using the likelihood function is not always made clear. Third, it is sometimes said that "missing completely at random" (MCAR) is needed for frequentist inference; at other times MAR is said to be sufficient.

While it is clear that some researchers writing on the theory of missing data have known what they intended, the omission of details by many authors, together with the seemingly different conditions assumed by different authors, make it difficult for readers to know precisely what was meant, and also to compare the work of different authors. This confusion has implications for statistical practice, since data analysts are encouraged to consider the plausibility of the MAR assumption before applying certain methods of analysis (e.g., [38]), but if the conscientious analyst consults the theoretical literature they will struggle to find a clear consensus on definitions and on how they relate to the validity of possible analytic approaches. Further confusion surrounds the concept of "ignorability", which does not seem to be well understood by practitioners and may be misinterpreted as providing a broad licence to ignore the fact that not all the desired data have been observed.

Shaun Seaman is Senior Statistician and Dan Jackson is Senior Statistician, MRC Biostatistics Unit, Cambridge, United Kingdom (e-mail: shaun.seaman@mrc-bsu.cam.ac.uk). John Galati is Senior Research Officer, Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute, and Department of Mathematics and Statistics, La Trobe University, Victoria, Australia. John Carlin is Director, Clinical Epidemiology and Biostatistics Unit, Murdoch Childrens Research Institute and University of Melbourne, Victoria, Australia.

In the present article, our objectives are to: (1) draw attention to the various gaps and inconsistencies in some definitions of MAR used in the literature; (2) provide unambiguous formulations of relevant MAR definitions; and (3) explain the relation between MAR and ignorability under different frameworks of statistical inference and, in so doing, identify the need for more than one definition of MAR.

The structure of the paper is as follows. In Section 2 we provide definitions of two distinct MAR conditions, one stronger than the other, and likewise for MCAR. The inconsistency in previous usage of the terms "MAR" and "MCAR" is documented in Section 3. The definitions of MAR and MCAR are central to the concept of ignorability, the definition of which varies according to the chosen framework of statistical inference. In Section 4 we distinguish between direct-likelihood inference, Bayesian inference, frequentist inference using the likelihood function and the frequentist properties of Bayesian estimators. Section 5 contains an explanation of which MAR/MCAR conditions are needed for the missingness mechanism to be ignorable for each of these types of inference. Section 6 covers the use of conditional likelihood and repeated sampling. We end with a discussion.

2. TWO DEFINITIONS OF MAR AND MCAR

We use $\mathbf{Y}$ to denote the vector of potentially observable data values (on all sample units), which for modelling purposes we treat as a random variable. Let $\mathbf{M}$ denote a vector of missingness indicators of the same length as $\mathbf{Y}$. The jth element of $\mathbf{M}$ equals one if the jth element of $\mathbf{Y}$ is observed and zero if it is missing. Let $o(\mathbf{Y}, \mathbf{M})$, a function of $\mathbf{Y}$ and $\mathbf{M}$, denote the subvector of $\mathbf{Y}$ consisting of elements whose corresponding elements of $\mathbf{M}$ equal one. So, $o(\mathbf{Y}, \mathbf{M})$ contains the observed elements of $\mathbf{Y}$. Let K denote the length of $o(\mathbf{Y}, \mathbf{M})$. So, K is a random variable and is equal to the sum of the elements of $\mathbf{M}$. When no elements of $\mathbf{Y}$ are observed, $o(\mathbf{Y}, \mathbf{M})$ is the empty set and $K = 0$. The reader may be familiar with the notation $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$. We choose not to use this notation because it is ambiguous, as we explain in Section 3. However, our notation $o(\mathbf{Y}, \mathbf{M})$ is equivalent to $\mathbf{Y}_{\text{obs}}$ as usually interpreted. When we consider a specific sample, it is convenient to have notation for the realised values of the random variables $\mathbf{M}$ and $\mathbf{Y}$; we denote these realised values as $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{y}}$, respectively. "Realised" and "observed" values should not be confused. The observed value, $o(\mathbf{Y}, \mathbf{M})$, of $\mathbf{Y}$ is a random variable and has a realised value, $o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}})$. The values of $\tilde{\mathbf{m}}$ and $o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}})$ are known, but that of $\tilde{\mathbf{y}}$ is only known if all elements of $\tilde{\mathbf{m}}$ equal one.

In the special case where the data are modelled as a set of J random variables measured on each of n units, as is often the case, $\mathbf{Y}$ is a vector of length nJ. Although in this special case one might alternatively define $\mathbf{Y}$ as a matrix with n rows and J columns, for the sake of generality we do not do this. For example, suppose that $\mathbf{Y}$ consists of two random variables, X and Z, measured on each of two units, that the realised value of (X, Z) is $(10, 3)$ for the first unit and $(4, 2)$ for the second, and that X is observed for both units but Z is only observed for the second. Then $\tilde{\mathbf{y}} = (10, 3, 4, 2)^T$, $\tilde{\mathbf{m}} = (1, 0, 1, 1)^T$ and $o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}}) = (10, 4, 2)^T$. Note that $o(\mathbf{y}, \mathbf{m})$ cannot be interpreted without the accompanying value of $\tilde{\mathbf{m}}$.

Consider a hypothesised "missingness model", that is, a model for the conditional distribution of $\mathbf{M}$ given $\mathbf{Y}$. Let $g_\phi(\mathbf{m} \mid \mathbf{y})$ denote the probability that $\mathbf{M} = \mathbf{m}$ given that $\mathbf{Y} = \mathbf{y}$ according to this model, where $\boldsymbol{\phi}$ is an unknown parameter. We now present two definitions of MAR.

DEFINITION 1. The data are realised MAR if $\forall \boldsymbol{\phi}$,

$$g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) = g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) \qquad \forall \mathbf{y} \text{ such that } o(\mathbf{y}, \tilde{\mathbf{m}}) = o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}})$$

(where $\mathbf{y}$ represents a value of $\mathbf{Y}$). This means that the hypothesised missingness model always (i.e., for all values of $\boldsymbol{\phi}$) assumes that the conditional probability that the missingness pattern $\mathbf{M}$ is its realised value $\tilde{\mathbf{m}}$, given the realised values of the elements of the data $\mathbf{Y}$ that are observed when $\mathbf{M} = \tilde{\mathbf{m}}$ and the values of the remaining, missing, elements, does not depend on these missing elements. Rubin [33] expressed this as follows: "The missing data are missing at random if for each possible value of the parameter $\boldsymbol{\phi}$, 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". There are several things to note about this definition. First, it is a statement only about the realised missingness pattern and realised observed data, not about missingness patterns or observed data that could have been realised but were not. Second, it is a statement about a hypothesised missingness model, rather than necessarily the true missingness process.

DEFINITION 2. The data are everywhere MAR if $\forall \boldsymbol{\phi}$,

$$g_\phi(\mathbf{m} \mid \mathbf{y}) = g_\phi(\mathbf{m} \mid \mathbf{y}^*) \qquad \forall \mathbf{m}, \mathbf{y}, \mathbf{y}^* \text{ such that } o(\mathbf{y}, \mathbf{m}) = o(\mathbf{y}^*, \mathbf{m})$$

(where $\mathbf{y}$ and $\mathbf{y}^*$ represent a pair of values of $\mathbf{Y}$). This means that the hypothesised missingness model always assumes that, for any value of the data, the probability of any possible missingness pattern, given the values of the corresponding observed elements and missing elements of the data, does not depend on the values of the missing elements. In order to make more obvious the difference between realised and everywhere MAR, note that Definition 1 can be rewritten as follows. The data are realised MAR if $\forall \boldsymbol{\phi}$, $g_{\phi}(\tilde{\mathbf{m}} \mid \mathbf{y}) = g_{\phi}(\tilde{\mathbf{m}} \mid \mathbf{y}^*)$ $\forall \mathbf{y}, \mathbf{y}^*$ such that $o(\mathbf{y}, \tilde{\mathbf{m}}) = o(\mathbf{y}^*, \tilde{\mathbf{m}}) = o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}})$. Unlike realised MAR, everywhere MAR is a statement about all possible missingness patterns and values of the observed data. Note that everywhere MAR implies realised MAR.

To illustrate and clarify the notation that we have used here, consider the example given above, that is, $\tilde{\mathbf{y}} = (10, 3, 4, 2)^T$, $\tilde{\mathbf{m}} = (1, 0, 1, 1)^T$ and $o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}}) = (10, 4, 2)^T$. The data are realised MAR if $\forall \boldsymbol{\phi}$, $g_{\phi}((1, 0, 1, 1)^T \mid \mathbf{y}) = g_{\phi}((1, 0, 1, 1)^T \mid \mathbf{y}^*)$ $\forall \mathbf{y}, \mathbf{y}^*$ such that the first, second and fourth elements of both $\mathbf{y}$ and $\mathbf{y}^*$ equal, respectively, 10, 4 and 2. That is, the data are realised MAR if, for any $\boldsymbol{\phi}$, $g_{\phi}((1, 0, 1, 1)^T \mid (10, a, 4, 2)^T) = g_{\phi}((1, 0, 1, 1)^T \mid (10, b, 4, 2)^T)$ for all a, b in the sample space of the second element of $\mathbf{Y}$.

Now consider the special case of independent identically distributed (i.i.d.) data, that is, $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_n^T)^T$ and $\mathbf{M} = (\mathbf{M}_1^T, \ldots, \mathbf{M}_n^T)^T$, where $(\mathbf{Y}_i, \mathbf{M}_i)$ $(i = 1, \ldots, n)$ are i.i.d. Let $o_1(\mathbf{Y}_i, \mathbf{M}_i)$ denote the subvector of $\mathbf{Y}_i$ consisting of elements of $\mathbf{Y}_i$ whose corresponding elements of $\mathbf{M}_i$ equal one. [Note that the function o_1 is analogous to the previously defined $o(\cdot)$, but whereas $o(\cdot)$ is a function of all the data, o_1 is a function of only the data for a single unit.] So, $\mathbf{Y}_i$, $\mathbf{M}_i$ and $o_1(\mathbf{Y}_i, \mathbf{M}_i)$ denote the data, the missingness pattern and the observed data, respectively, for the ith of n units. Consider a hypothesised model for the conditional distribution of $\mathbf{M}_i$ given $\mathbf{Y}_i$, and let $g_{\phi,1}(\mathbf{m}_i \mid \mathbf{y}_i)$ denote the probability that $\mathbf{M}_i = \mathbf{m}_i$ given that $\mathbf{Y}_i = \mathbf{y}_i$ according to the model. In this case, Definitions 2 and 3 are equivalent.

DEFINITION 3. The data are everywhere MAR if $\forall i, \boldsymbol{\phi}$,

$$g_{\phi,1}(\mathbf{m}_i \mid \mathbf{y}_i) = g_{\phi,1}(\mathbf{m}_i \mid \mathbf{y}_i^*) \quad \forall \mathbf{y}_i, \mathbf{y}_i^* \text{ such that } o_1(\mathbf{y}_i, \mathbf{m}_i) = o_1(\mathbf{y}_i^*, \mathbf{m}_i).$$

Definition 3 may only be applied when $(\mathbf{Y}_1, \mathbf{M}_1), \ldots, (\mathbf{Y}_n, \mathbf{M}_n)$ are i.i.d. If, for example, $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ were i.i.d. and $\mathbf{M}_1, \ldots, \mathbf{M}_n$ were identically distributed but with $\mathbf{M}_i$ depending on $\mathbf{M}_j$ and/or $\mathbf{Y}_j$ for $j \neq i$, then $(\mathbf{Y}_1, \mathbf{M}_1), \ldots, (\mathbf{Y}_n, \mathbf{M}_n)$ would not be i.i.d. and so Definition 3 could not apply. The data might nevertheless still be everywhere MAR by Definition 2.

Finally, we present two definitions of MCAR.

DEFINITION 4. The data are realised MCAR if $\forall \boldsymbol{\phi}$,

$$g_{\phi}(\tilde{\mathbf{m}} \mid \mathbf{y}) = g_{\phi}(\tilde{\mathbf{m}} \mid \mathbf{y}^*) \quad \forall \mathbf{y}, \mathbf{y}^*.$$

DEFINITION 5. The data are everywhere MCAR if $\forall \boldsymbol{\phi}$,

$$g_{\phi}(\mathbf{m} \mid \mathbf{y}) = g_{\phi}(\mathbf{m} \mid \mathbf{y}^*) \quad \forall \mathbf{m}, \mathbf{y}, \mathbf{y}^*.$$

Realised MCAR means that the probability of the realised missingness pattern given the data does not depend on the data. Realised MCAR implies realised MAR but not everywhere MAR. Everywhere MCAR means that the probability of *any* missingness pattern given the data does not depend on the data, that is, $\mathbf{M}$ is independent of $\mathbf{Y}$. Everywhere MCAR implies realised MCAR, realised MAR and everywhere MAR.

3. MAR AND MCAR IN THE LITERATURE: A REVIEW

Historically, the first definition of MAR was that of Rubin (1976) [33]. This is Definition 1, that is, the definition for realised MAR (apart from minor differences in notation and the fact that Rubin's definition begins "The missing data are MAR" rather than "The data are MAR"). Rubin (1987) [36] largely avoided the term "MAR", preferring instead the terms "ignorable sampling" and "ignorable response". However, he did (page 53) briefly discuss the relation between these three terms. It is evident from that discussion that he was using the Rubin (1976) [33] definition. Heitjan and colleagues, in a series of papers (e.g., [10–15]), consistently used "MAR" to mean realised MAR. Harel and Schafer [9] also defined realised MAR. Most other authors have used "MAR" to mean everywhere MAR.

Several authors (Schafer [37]; Kenward and Molenberghs [20]; Lu and Copas [25]; Jaeger [17]) provided definitions of everywhere MAR but accompanied this definition with a citation of Rubin (1976) [33] (which defines realised, rather than everywhere, MAR). In fact, most of these authors said explicitly that their definition was an expression of Rubin's (1976) [33] definition.

The potential of the variety of definitions of MAR to cause confusion was illustrated by an exchange of letters between Heitjan [13] and Diggle [5]. Note that according to Rubin's (1976) [33] definition (i.e., realised

MAR), if all the data are observed, they cannot fail to be MAR (although one might alternatively say that his definition is a statement about *the missing data* and in this situation there are no missing data, so there are no missing data to be MAR). Heitjan gave an example in which a single variable X is measured on n individuals and could potentially be missing on some of these individuals. However, he supposed that in the data set actually observed, X is observed on all n individuals, so there are no missing data. He stated that the data are MAR. Diggle responded by saying that the data are not MAR, since the probability that X is observed depends on X, which could be missing. The reason for this disagreement is that Heitjan was using the definition of realised MAR whereas Diggle was using that of everywhere MAR.

In addition to the problems caused by this dual use of the term "MAR", definitions of MAR found in some of the key literature on missing data, including textbooks, contain certain ambiguities.

Many authors (Little and Rubin [23, 24]; Schafer [37]; Kenward and Molenberghs [20]; Harel and Schafer [9]; Fitzmaurice et al. [8]) used the problematic notation $\mathbf{Y}_{\text{obs}}$ and $\mathbf{Y}_{\text{mis}}$ mentioned in Section 2. Little and Rubin [23, 24], for example, said that $\mathbf{Y}_{\text{obs}}$ denotes the observed components or entries of $\mathbf{Y}$, that $\mathbf{Y}_{\text{mis}}$ denotes the missing components, and that the missing data mechanism is called MAR if

$$(1)\qquad f(\mathbf{M} \mid \mathbf{Y}, \boldsymbol{\phi}) = f(\mathbf{M} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\phi}) \quad \forall \mathbf{Y}_{\text{mis}}, \boldsymbol{\phi}$$

[where $f(\cdot \mid \cdot)$ denotes a conditional distribution]. The notation $f(\mathbf{M} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\phi})$ is somewhat confusing, because $\mathbf{Y}_{\text{obs}}$ is itself a function of $\mathbf{M}$. Interpreted literally, $\mathbf{Y}_{\text{obs}} = o(\mathbf{Y}, \mathbf{M})$. Hence, if $\mathbf{Y}_{\text{obs}}$ is known, then K is also known, and so $f(\mathbf{M} \mid \mathbf{Y}_{\text{obs}}, \boldsymbol{\phi})$ should equal zero unless the number of nonzero elements of $\mathbf{M}$ equals K. Nevertheless, we presume that equation (1) was intended to mean Definition 2 (i.e., everywhere MAR). Fitzmaurice et al. [8] gave a definition similar to equation (1), but added that this means $\mathbf{M}$ is conditionally independent of $\mathbf{Y}_{\text{mis}}$ given $\mathbf{Y}_{\text{obs}}$. This is rather difficult to interpret, given that $\mathbf{Y}_{\text{mis}}$ is a function of $\mathbf{M}$.

Another source of ambiguity concerns the parameter $\boldsymbol{\phi}$. Definitions 1–3 require a particular equality to hold for all values of $\boldsymbol{\phi}$. Several authors (Robins and Gill [31]; Kenward and Molenberghs [20]; Tsiatis [39]; Fitzmaurice et al. [8]) omitted the parameter $\boldsymbol{\phi}$ when defining MAR, with the result that it is not obvious whether equality is required to hold for all $\boldsymbol{\phi}$ or just for its "true" value. Schafer [37] did include $\boldsymbol{\phi}$, but was also unclear about whether equality must hold for all $\boldsymbol{\phi}$. Judging from the use that these authors made of their MAR assumptions, most of them seem implicitly to have meant that the equality should hold for all $\boldsymbol{\phi}$. However, Fitzmaurice et al. [8] seem to require equation (1) to hold only for the true value of $\boldsymbol{\phi}$: they appear to be referring to the "true" missingness mechanism, rather than to a model for the missingness. We shall return to this point in Section 7.

Just as there can be ambiguity about $\boldsymbol{\phi}$, it is sometimes not entirely clear whether a definition of MAR requires an equality to hold for all $\mathbf{Y}$ or just for $\mathbf{Y}$ compatible with $o(\tilde{\mathbf{y}}, M)$. See, in particular, equation (1).

We have concentrated on MAR, but there is also ambiguity about the definition of MCAR. In his original 1976 paper [33], Rubin did not mention MCAR. He instead introduced the concept of the observed data being "observed at random". The realised MCAR definition (Definition 4) is equivalent to the combination of the missing data being realised MAR and the observed data being observed at random [11] (see also Little [21]). Heitjan and colleagues have used "MCAR" to mean realised MCAR. Many other authors (e.g., Little and Rubin [23, 24] and [37]) have used "MCAR" to mean everywhere MCAR. In the situation of repeated-measures outcome data with fully observed covariates, Molenberghs and Kenward [26] used "MCAR" to mean that missingness in the outcomes cannot depend on the outcomes but can depend on the covariates. Elsewhere this has been called "covariate-dependent MCAR" [22, 41].

4. DIRECT-LIKELIHOOD, BAYESIAN AND FREQUENTIST INFERENCE

In Section 5 we shall discuss ignorability. The definition of ignorability depends on the framework of inference adopted. Here we review the distinctions between four types of inference: Bayesian inference, direct-likelihood inference (also known as pure-likelihood inference), general frequentist inference and frequentist likelihood inference. For simplicity of exposition, we describe inference when the data $\mathbf{Y}$ are fully observed. In Section 5 we describe the generalisation to incomplete data.

In Bayesian and direct-likelihood inference a probability distribution function is specified for the data $\mathbf{Y}$. This function involves a finite set of unknown parameters, $\boldsymbol{\theta}$. Some of these are of interest and the aim is to make inference about their values; others may be nuisance parameters. The likelihood is defined as any multiple of this probability distribution function where the multiplier does not depend on any of the parameters. Whereas the probability distribution function is regarded as a function of the data with the values of the parameters considered fixed, the likelihood is regarded as a function of the parameters with the data considered fixed.

In direct-likelihood inference [2, 28, 30], the value of the parameters at which the likelihood is a maximum (the maximum likelihood estimate) is used as a point estimate and the ratio of the value of the likelihood at different parameter values is used to judge which parameter values are plausible. The normalised likelihood is defined as the likelihood divided by the value of the likelihood at the maximum likelihood estimate (so that the normalised likelihood takes value one at the maximum likelihood estimate). When there is only one parameter, a likelihood interval is defined as the set of parameter values within which the values of the normalised likelihood are greater than some threshold. Different thresholds have been proposed, for example, Fisher [7] suggested $1/15$ and Royall [32] suggested $1/32$.

When there is more than one parameter, a likelihood interval for any one of them can be obtained by first eliminating the others. Two commonly used ways to eliminate parameters are the profile likelihood method and the conditional likelihood method. Suppose, without loss of generality, that $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_2$ are the parameters to be eliminated. The profile likelihood for θ_1 is defined as the function obtained, for each possible value of θ_1, by fixing θ_1 at that value and then maximising the likelihood for $\boldsymbol{\theta}$ over the space of $\boldsymbol{\theta}_2$. In the profile likelihood method, a likelihood interval for θ_1 is calculated using the profile likelihood for θ_1 in place of the likelihood for $\boldsymbol{\theta}$. In the conditional likelihood method, a conditional probability distribution function is specified for $\mathbf{Y}$ given a (possibly vector) function of $\mathbf{Y}$. The resulting conditional likelihood contains fewer parameters than the unconditional likelihood, that is, that based on the unconditional probability distribution function for $\mathbf{Y}$. If the conditional likelihood contains only θ_1, it can be used to construct a likelihood interval for θ_1. If it contains additional parameters, these can be eliminated using the profile likelihood method. There is no clear theoretical basis for choosing between the profile likelihood and conditional likelihood approaches, and each appear to have their merits for different situations.

In Bayesian inference, uncertainty about parameters is represented directly by probability models, requiring a prior distribution to be specified. The posterior distribution of the parameters is obtained by Bayes' theorem. For any of the parameters in the model, the mean of its posterior distribution is typically used as a point estimate and (α_l, α_u) used as an interval of uncertainty (a credible interval), where α_l and α_u are the lth and uth centiles (e.g., 2.5th and 97.5th) of that parameter's marginal posterior distribution. This interval is interpretable as meaning that the posterior probability that the parameter lies within (α_l, α_u) is $(u - l)/100$. The use of the marginal posterior distribution means that all other parameters are eliminated by integrating them out of the joint posterior distribution of all the parameters.

In direct-likelihood inference and Bayesian inference as described above, only the realised value of $\mathbf{Y}$ is of interest; there is no consideration of other values of $\mathbf{Y}$ that could have been realised but which were not. Frequentist inference, on the other hand, is concerned with the (hypothetical) repeated sampling of $\mathbf{Y}$ and with the properties of inferential summaries such as point and interval estimates under this repeated sampling. It is only when repeated sampling is considered that the concepts of bias, standard error, efficiency, power and confidence interval become meaningful. The bias of an estimator of a parameter, for example, is defined as the difference between the mean of the sampling distribution of the estimator and the true value of the parameter; the standard error is the standard deviation of the sampling distribution of the estimator; a confidence interval is an interval obtained using a rule that has a stated probability of producing an interval containing the true value of the parameter in a repeated sample. One important example of a rule for constructing confidence intervals is the rule used in direct-likelihood inference to construct likelihood intervals, that is, a likelihood interval becomes, in the framework of frequentist inference, a confidence interval.

In frequentist inference, a function $s(\mathbf{Y})$ of $\mathbf{Y}$ is chosen and its realised value, $s(\tilde{\mathbf{y}})$, is compared with the sampling distribution of $s(\mathbf{Y})$, that is, the distribution of $s(\mathbf{Y})$ in repeated samples. This sampling distribution may be conditional on the realised value of a (possibly vector) function of $\mathbf{Y}$. We distinguish between general frequentist inference, where $s(\mathbf{Y})$ can be any function of $\mathbf{Y}$, and frequentist likelihood inference, where $s(\mathbf{Y})$ depends on $\mathbf{Y}$ only through the likelihood of $\mathbf{Y}$. Frequentist likelihood inference includes using the observed or expected information to estimate the standard error of the maximum likelihood estimator (MLE), using this MLE and standard error to construct a confidence interval, using likelihood intervals as confidence intervals, and using likelihood-ratio, Wald and score tests. Frequentist likelihood inference is like direct-likelihood inference, in that it also uses the MLE and likelihood intervals, but goes beyond it, in that it involves claims about the behaviour of the MLE and likelihood intervals in repeated samples. Frequentist likelihood inference is often referred to simply as "likelihood inference".

Often even statisticians using Bayesian methods are interested in frequentist properties of their estimators, for example, the bias of the posterior mean or the coverage of a credible interval [19, 35].

The distinction between direct-likelihood inference and frequentist likelihood inference has not always been made clear in the literature. For example, Heitjan and Rubin [15] and Harel and Schafer [9] referred to direct-likelihood inference simply as "likelihood inference". Molenberghs et al. [27] appear to use the term "direct-likelihood analysis" when writing about repeated sample properties of the likelihood. Also, the potential interest in frequentist properties of Bayesian estimators has rarely been mentioned in the literature on missing data, except in the context of multiple imputation.

5. IGNORABILITY OF THE MISSINGNESS MECHANISM

In this section we clarify which missingness assumption suffices for the missingness mechanism to be ignorable for each of the types of inferences described in Section 4. Intuitively, "ignorable" means that inferences obtained from a parametric model for the data alone are the same as inferences obtained from a joint model for the data and missingness mechanism. To serve as a workable definition, one needs to say what is meant by "the same", and in the literature authors have not always been explicit on this point. We endeavour to be clear, but defer specification of our definitions to the relevant subsections below.

Consider a joint parametric model for the complete data $\mathbf{Y}$ and missingness pattern $\mathbf{M}$. Let $f_\theta(\mathbf{y})g_\phi(\mathbf{m} \mid \mathbf{y})$ denote the joint distribution of $\mathbf{Y}$ and $\mathbf{M}$ according to this model, and let $\Omega_{\theta,\phi}$ denote the joint parameter space for $(\boldsymbol{\theta}, \boldsymbol{\phi})$. Let $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{m}}$ be a given realisation of $\mathbf{Y}$ and $\mathbf{M}$. Let $\Omega_\theta = \pi_1(\Omega_{\theta,\phi})$ and $\Omega_\phi = \pi_2(\Omega_{\theta,\phi})$ be the parameter spaces for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively, corresponding to the joint parameter space $\Omega_{\theta,\phi}$. Following Heitjan and Basu [14], we avoid measure-theoretic difficulties by assuming that $\mathbf{Y}$ is discrete. Because in reality all data are measured to finite precision, this assumption is not restrictive. Reference to continuous distributions should be interpreted as meaning discrete distributions on a fine grid, and integrals can be interpreted as sums.

The *joint likelihood for* $(\boldsymbol{\theta}, \boldsymbol{\phi})$ is the function with domain $\Omega_{\theta,\phi}$ given by

$$(2) \quad L_1(\boldsymbol{\theta}, \boldsymbol{\phi}) = \int f_\theta(\mathbf{y})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y})r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\,d\mathbf{y},$$

where $r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})$ equals one if $o(\mathbf{y}, \tilde{\mathbf{m}}) = o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}})$ and zero otherwise. Note that the integral here integrates out the missing data. The *likelihood for* $\boldsymbol{\theta}$ *ignoring the missing-data mechanism* is the function with domain Ω_θ given by

$$(3) \qquad L_2(\boldsymbol{\theta}) = \int f_\theta(\mathbf{y})r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\,d\mathbf{y}.$$

For any fixed $\boldsymbol{\phi} \in \Omega_\phi$, the *fixed-*$\boldsymbol{\phi}$ *likelihood for* $\boldsymbol{\theta}$ is the function with domain Ω_θ given by

$$(4) \quad \begin{aligned} L_{3,\phi}(\boldsymbol{\theta}) &= \delta\{(\boldsymbol{\theta}, \boldsymbol{\phi}), \Omega_{\theta,\phi}\} \\ &\quad \cdot \int f_\theta(\mathbf{y})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y})r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\,d\mathbf{y}, \end{aligned}$$

where $\delta\{(\boldsymbol{\theta}, \boldsymbol{\phi}), \Omega_{\theta,\phi}\}$ equals one if $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Omega_{\theta,\phi}$ and zero otherwise. The *profile likelihood for* $\boldsymbol{\theta}$ is the function with domain Ω_θ given by

$$(5) \quad \begin{aligned} L_4(\boldsymbol{\theta}) = \max_{\boldsymbol{\phi} \in \Omega_\phi} \biggl[&\delta\{(\boldsymbol{\theta}, \boldsymbol{\phi}), \Omega_{\theta,\phi}\} \\ &\cdot \int f_\theta(\mathbf{y})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y})r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\,d\mathbf{y} \biggr]. \end{aligned}$$

In Section 6 we shall consider the use of conditional likelihoods.

5.1 Direct-Likelihood Inference

The main work on ignorability for direct-likelihood inference can be summed up in the following theorem. After giving a proof, we shall discuss why this theorem has been considered to justify the use of L_2, the likelihood for $\boldsymbol{\theta}$ ignoring the missing-data mechanism, when the data are realised MAR and the parameters are distinct.

THEOREM 1. *If realised MAR holds and* $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$, *then*: (i) $L_1(\boldsymbol{\theta}, \boldsymbol{\phi})$ *factorises into two components, such that each parameter appears in only one component*; (ii) *for any* $\boldsymbol{\phi} \in \Omega_\phi$ *satisfying* $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) > 0$, $L_{3,\phi}(\boldsymbol{\theta})$ *is proportional to* $L_2(\boldsymbol{\theta})$; *and* (iii) *if* $\exists \boldsymbol{\phi} \in \Omega_\phi$ *such that* $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) > 0$, *then* $L_4(\boldsymbol{\theta})$ *is a special case of* $L_{3,\phi}(\boldsymbol{\theta})$ *and, hence,* $L_4(\boldsymbol{\theta})$ *is proportional to* $L_2(\boldsymbol{\theta})$.

PROOF. As $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$, it follows that whenever $\boldsymbol{\phi} \in \Omega_\phi$ and $\boldsymbol{\theta} \in \Omega_\theta$, then $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Omega_{\theta,\phi}$, and so $\delta\{(\boldsymbol{\theta}, \boldsymbol{\phi}), \Omega_{\theta,\phi}\} = 1$. So, for $\boldsymbol{\phi} \in \Omega_\phi$ and $\boldsymbol{\theta} \in \Omega_\theta$,

$$L_1(\boldsymbol{\theta}, \phi) = \int f_\theta(\mathbf{y}) g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\, d\mathbf{y} \tag{6}$$

$$= g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) \int f_\theta(\mathbf{y}) r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\, d\mathbf{y} \tag{7}$$

$$= L_5(\boldsymbol{\phi}) L_2(\boldsymbol{\theta}), \tag{8}$$

where

$$L_5(\boldsymbol{\phi}) = g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}})$$

is a function of $\boldsymbol{\phi}$ only. Hence, (i) is true. Note that line (7) follows because of realised MAR.

If realised MAR holds and $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$, line (7) is equal to $L_{3,\phi}(\boldsymbol{\theta})$. Since $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}})$ is not a function of $\boldsymbol{\theta}$, it then follows that $L_{3,\phi}(\boldsymbol{\theta})$ is proportional to $L_2(\boldsymbol{\theta})$ for any $\boldsymbol{\phi} \in \Omega_\phi$ such that $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) > 0$. So, (ii) is true.

Likewise, when the data are realised MAR and $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$,

$$L_4(\boldsymbol{\theta}) = \int f_\theta(\mathbf{y}) r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\, d\mathbf{y} \times \max_{\boldsymbol{\phi} \in \Omega_\phi} g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}).$$

The function $\max_{\boldsymbol{\phi} \in \Omega_\phi} g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}})$ does not depend on $\boldsymbol{\theta}$. Moreover, it is nonzero when $\exists \boldsymbol{\phi} \in \Omega_\phi$ such that $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) > 0$. So, $L_4(\boldsymbol{\theta}) = L_{3,\hat{\phi}}(\boldsymbol{\theta})$, where $\hat{\boldsymbol{\phi}}$ is the value of $\boldsymbol{\phi}$ that maximises $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}})$. Hence, (iii) is true. □

In the literature, this factorisation of the joint likelihood and this proportionality of likelihoods have been used as a basis for defining when the missingness mechanism can be ignored when performing direct-likelihood inference. Rubin [33], for example, used the proportionality of likelihoods to write: "When making direct-likelihood or Bayesian inferences about $\boldsymbol{\theta}$, 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 $\boldsymbol{\theta}$". Anscombe [1] wrote that when the joint likelihood for a parameter of interest $\boldsymbol{\theta}$ and a nuisance parameter $\boldsymbol{\phi}$ factorises into two components, such that each parameter appears in only one component, information on each factor can be considered separately. The same was written by Hinde and Aitkin [16]. Royall [32] called the component involving $\boldsymbol{\theta}$ the "likelihood for $\boldsymbol{\theta}$" and said that the relative support for any two values of $\boldsymbol{\theta}$ is given by the ratio of the values of this likelihood evaluated at those two $\boldsymbol{\theta}$ values. Edwards [6] supported the use of the profile likelihood when the joint likelihood factorises. He wrote: "since the value of $\boldsymbol{\phi}$ is irrelevant to our inference on $\boldsymbol{\theta}$, replacing $\boldsymbol{\phi}$ in [the joint likelihood] by its maximum likelihood estimate will not invalidate the likelihood". Kalbfleisch and Sprott [18] agreed with Edwards. When comparing inference for $\boldsymbol{\theta}$ using L_1 and L_2 in situations where the two may give different answers, Heitjan [10, 15], pages 1103 and 2249, interpreted inference for $\boldsymbol{\theta}$ using L_1 as meaning inference using the profile likelihood. Tsou and Royall [40] considered the strength of evidence in the presence of a nuisance parameter as being the strength of evidence that would be in the data if the value of the nuisance parameter were known. That is, they considered the strength of evidence to be the particular fixed-$\boldsymbol{\phi}$ likelihood for $\boldsymbol{\theta}$ corresponding to the true value of $\boldsymbol{\phi}$.

All these authors, therefore, provide justification for interpreting Theorem 1 as meaning that direct-likelihood inference about $\boldsymbol{\theta}$ can be performed using L_2 when the data are realised MAR and $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct parameters.

To picture the effect of realised MAR and distinctness of parameters on the joint likelihood $L_1(\boldsymbol{\theta}, \boldsymbol{\phi})$, it is helpful to consider a joint model where $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are both scalar parameters. The graph of L_1 is then a surface in three dimensions lying above a (θ, ϕ) plane. The realised MAR condition imposes geometric structure on this surface [evident from equations (7) and (8)] such that curves obtained from the surface by fixing ϕ at various values are all proportional, simply being copies of L_2 scaled by the conditional probability of realising the observed missingness pattern under the given $\boldsymbol{\phi}$ value. The function L_1 is, however, only defined for values of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ in $\Omega_{\theta,\phi}$. Hence, the curve formed from the L_1 by fixing $\boldsymbol{\phi}$ may be undefined for some values of $\boldsymbol{\theta}$ where the L_2 curve is defined. So, one can think of each curve formed from L_1 by fixing $\boldsymbol{\phi}$ as being a proportional copy of L_2 with, potentially, one or more sections omitted. The assumption of distinct parameters ensures that such "omitted" sections do not exist, and therefore that the curves are proportional at all $\boldsymbol{\theta}$ values in Ω_θ.

So far, we have considered the elimination of $\boldsymbol{\phi}$ as a nuisance parameter. As discussed in Section 4, when a likelihood interval is required for a single element, θ_1, of a vector parameter, $\boldsymbol{\theta}$, the other parameters, $\boldsymbol{\theta}_2$, are also nuisance parameters and must be eliminated. If $\boldsymbol{\theta}_2$ is eliminated from $L_2(\boldsymbol{\theta})$ and $L_4(\boldsymbol{\theta})$ using the profile likelihood method, the proportionality of $L_4(\boldsymbol{\theta})$ and $L_2(\boldsymbol{\theta})$ also ensures the proportionality of the profile likelihoods for θ_1 derived from $L_2(\boldsymbol{\theta})$ and $L_4(\boldsymbol{\theta})$.

Hence, the likelihood intervals for θ_1 obtained from L_2 and L_4 will be the same. We discuss the use of conditional likelihood in Section 6.

5.2 Bayesian Inference

Consider Bayesian inference accounting for the missingness mechanism. Let $p_{\theta,\phi}(\boldsymbol{\theta}, \boldsymbol{\phi})$ denote the joint prior distribution of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and let $p_\theta(\boldsymbol{\theta})$ denote the corresponding marginal prior distribution of $\boldsymbol{\theta}$. The missingness mechanism is said to be ignorable for Bayesian inference if the marginal posterior distribution of $\boldsymbol{\theta}$ obtained from modelling both the complete data, **Y**, and the missingness pattern, **M**, is equal to the posterior for $\boldsymbol{\theta}$ obtained by modelling **Y** alone. The main work in this area can be summed up by the following theorem.

THEOREM 2. *Suppose that* (1) *the data are realised MAR and* (2) $\boldsymbol{\theta}$ *and* $\boldsymbol{\phi}$ *are a priori independent. The posterior distribution of* $\boldsymbol{\theta}$ *that results from using the likelihood* $L_2(\boldsymbol{\theta})$ *and the prior* $p(\boldsymbol{\theta})$ *is the same as the posterior distribution that results from using likelihood* $L_1(\boldsymbol{\theta}, \boldsymbol{\phi})$ *and prior* $p_{\theta,\phi}(\boldsymbol{\theta}, \boldsymbol{\phi})$.

PROOF. When $L_1(\boldsymbol{\theta}, \boldsymbol{\phi})$ and $p_{\theta,\phi}(\boldsymbol{\theta}, \boldsymbol{\phi})$ are used, the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ is proportional to $p_{\theta,\phi}(\boldsymbol{\theta}, \boldsymbol{\phi})L_1(\boldsymbol{\theta}, \boldsymbol{\phi})$. If $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are a priori independent, $p_{\theta,\phi}(\boldsymbol{\theta}, \boldsymbol{\phi})$ factorises as $p_\theta(\boldsymbol{\theta})p_\phi(\boldsymbol{\phi})$, where $p_\phi(\boldsymbol{\phi})$ is the marginal prior for $\boldsymbol{\phi}$. If, furthermore, the data are realised MAR, it follows from equation (8) that the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\phi})$ is proportional to $p_\phi(\boldsymbol{\phi})L_5(\boldsymbol{\phi})p_\theta(\boldsymbol{\theta})L_2(\boldsymbol{\theta})$. Since $p_\phi(\boldsymbol{\phi})L_5(\boldsymbol{\phi})$ is a function of $\boldsymbol{\phi}$ only, the marginal posterior distribution of $\boldsymbol{\theta}$ is proportional to $p_\theta(\boldsymbol{\theta})L_2(\boldsymbol{\theta})$. This is the same posterior distribution that is obtained if $L_2(\boldsymbol{\theta})$ and $p_\theta(\boldsymbol{\theta})$ are used. □

5.3 General Frequentist Inference

From the joint model, for any $\boldsymbol{\phi} \in \Omega_\phi$ for which $\exists\mathbf{y}$ such that $f_\theta(\mathbf{y})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) > 0$, the conditional distribution of $o(\mathbf{Y}, \mathbf{M})$ given $\mathbf{M} = \tilde{\mathbf{m}}$ is

$$(9) \qquad \int f_\theta(\mathbf{u})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{u})r(\mathbf{u}, \mathbf{y}, \tilde{\mathbf{m}})\,d\mathbf{u} \Big/ \int f_\theta(\mathbf{u})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{u})\,d\mathbf{u}.$$

In general, this distribution may depend on $\boldsymbol{\phi}$. Let $t\{o(\mathbf{Y}, \mathbf{M}), \mathbf{M}\}$ be a function of $o(\mathbf{Y}, \mathbf{M})$ and **M**. Rubin [33] called the distribution of $t\{o(\mathbf{Y}, \mathbf{M}), \mathbf{M}\}$ given $\mathbf{M} = \tilde{\mathbf{m}}$ implied by the distribution of $o(\mathbf{Y}, \mathbf{M})$ given $\mathbf{M} = \tilde{\mathbf{m}}$ in expression (9) the "correct conditional sampling distribution" of $t\{o(\mathbf{Y}, \mathbf{M}), \mathbf{M}\}$. In general, the distribution given by (9) is not equal to

$$(10) \qquad \int f_\theta(\mathbf{u})r(\mathbf{u}, \mathbf{y}, \tilde{\mathbf{m}})\,d\mathbf{u}$$

and so, in general, the conditional distribution of $o(\mathbf{Y}, \mathbf{M})$ given $\mathbf{M} = \tilde{\mathbf{m}}$ is not that given by expression (10). Nevertheless, the latter distribution is the distribution that corresponds to likelihood $L_2(\boldsymbol{\theta})$. Heitjan and Basu [14] called the distribution of $t\{o(\mathbf{Y}, \mathbf{M}), \mathbf{M}\}$ given $\mathbf{M} = \tilde{\mathbf{m}}$ implied by the distribution in (10) the "potentially incorrect sampling distribution" of $t\{o(\mathbf{Y}, \mathbf{M}), \mathbf{M}\}$.

THEOREM 3. *When the data are realised MCAR and* $\exists\mathbf{y}$ *such that* $f_\theta(\mathbf{y})g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) > 0$, *the potentially incorrect sampling distribution is equal to the correct conditional sampling distribution.*

PROOF. If the data are realised MCAR, then for each value of $\boldsymbol{\phi}$ the value of $g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y})$ does not depend on **y**. Hence, expression (9) reduces to expression (10). □

Note that in Theorem 3 repeated sampling is conditional on the realised missingness pattern, that is, conditional on $\mathbf{M} = \tilde{\mathbf{m}}$. Little [21] argued that it is wrong to condition on $\mathbf{M} = \tilde{\mathbf{m}}$, as **M** is not an ancillary statistic for $\boldsymbol{\theta}$ unless the stronger condition of everywhere MCAR is satisfied. Rubin [34] disagreed, saying that "the usual definition of ancillary (Cox and Hinkley [3], page 35) is incorrect for inference about $\boldsymbol{\theta}$ and should be modified to be conditional on the observed value of the ancillary statistic". Heitjan [12] continued this discussion, introducing the concept of an observed ancillary statistic and agreeing with Rubin's assertion that the missingness pattern can be conditioned upon when the data are realised MCAR. As Rubin noted, although Theorem 3 might be regarded as a statement about when the missingness mechanism can be ignored, the realised missingness pattern itself is not ignored, because the repeated sampling is conditional on it.

As mentioned in Section 4, repeated sampling may be conditional on a function of **Y**. We discuss this in Section 6.

5.4 Frequentist Likelihood Inference

Since frequentist likelihood inference is a special case of general frequentist inference, Theorem 3 still applies. However, for frequentist likelihood inference a further result can be obtained when the data are everywhere MAR and $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct. When the data are not realised MCAR, $\mathbf{M}$ is not observed ancillary, and so repeated sampling should not be conditional on $\mathbf{M}$. However, if the data are everywhere MAR and $\Omega_{\boldsymbol{\theta},\boldsymbol{\phi}} = \Omega_{\boldsymbol{\theta}} \times \Omega_{\boldsymbol{\phi}}$, it follows from Theorem 1 that $L_2(\boldsymbol{\theta})$, $L_{3,\phi}(\boldsymbol{\theta})$ and $L_4(\boldsymbol{\theta})$ are proportional not only in the realised sample but also in repeated samples. Therefore, the MLE of $\boldsymbol{\theta}$, the estimated variance of this MLE calculated from the observed information matrix, likelihood intervals for $\boldsymbol{\theta}$, and likelihood-ratio, Wald and score test statistics for hypotheses concerning $\boldsymbol{\theta}$ will be the same in both the realised and repeated samples whether calculated using L_2 or L_1. That is, the same frequentist likelihood inference for $\boldsymbol{\theta}$ will be made whether one uses L_2 or L_1.

A similar result applies for Bayesian point estimators and credible intervals. Suppose that the data are everywhere MAR and, for every possible data vector $\mathbf{Y}$ and missingness pattern $\mathbf{M}$, the prior for $(\boldsymbol{\theta}, \boldsymbol{\phi})$ in the joint model can be written as $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = p(\boldsymbol{\theta}) \times p(\boldsymbol{\phi})$, where $p(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$ in the model that ignores the missingness pattern. Then, for every possible $(\mathbf{Y}, \mathbf{M})$, the posterior distribution for $\boldsymbol{\theta}$ derived from the likelihood L_1 and prior $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ of the joint model is the same as that derived from the likelihood L_2 and prior $p(\boldsymbol{\theta})$ of the model that ignores the missingness pattern. Consequently, under these assumptions, the repeated-sampling properties of Bayesian point estimators and credible intervals for $\boldsymbol{\theta}$ in repeated samples will be the same whether one uses L_1 and $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ and integrates over $\boldsymbol{\phi}$ or one uses L_2 and $p(\boldsymbol{\theta})$.

One important caveat needs to be stated. Standard errors can, in general, be calculated using either the expected or the observed information. When the data are everywhere MAR and $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct, the expected information from L_2 should not be used naively [20]. Using this expected information is only appropriate under the stronger assumption that the data are everywhere MCAR. It is recommended that the observed information be used instead [20].

6. CONDITIONAL LIKELIHOOD AND REPEATED SAMPLING

We now consider (1) conditional likelihoods and (2) repeated sampling conditional not only on $\mathbf{M} = \tilde{\mathbf{m}}$ but also on some function of $\mathbf{Y}$. Let $\mathbf{X} = b(\mathbf{Y})$ denote the function of $\mathbf{Y}$ being conditioned on and $\tilde{\mathbf{x}}$ denote the realised value of $\mathbf{X}$.

First, consider the use of conditional likelihood. One example of the use of a conditional likelihood is where data $\mathbf{Y}$ consist of a set of covariates and an outcome for a sample of individuals and this outcome is regressed on the covariates. When the covariates are fully observed, there is no need to specify a likelihood for all of $\mathbf{Y}$; instead, a likelihood for the outcomes conditional on the covariates is sufficient. Here, $\mathbf{X}$ consists of the covariates. A second example is conditional logistic regression for matched case-control data, where the likelihood is conditional on the number of cases and controls in each matched set. So here, $\mathbf{X}$ consists of these numbers of cases and controls.

Assume that either $\tilde{\mathbf{x}}$ is observed or $\int f_\theta(\mathbf{y} \mid \mathbf{x} = \tilde{\mathbf{x}}) r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\, d\mathbf{y}$ does not depend on the value of the missing part of $\tilde{\mathbf{x}}$. In equations (2)–(5), $f_\theta(\mathbf{y})$ should be replaced by $f_\theta(\mathbf{y} \mid \mathbf{x} = \tilde{\mathbf{x}})$, the conditional probability distribution of $\mathbf{Y}$ given $\mathbf{X} = \tilde{\mathbf{x}}$. Theorem 1 then still applies. Moreover, if the data are everywhere MAR, then $L_2(\boldsymbol{\theta})$ and $L_4(\boldsymbol{\theta})$ [both with $f_\theta(\mathbf{y})$ replaced by $f_\theta(\mathbf{y} \mid \mathbf{x} = \tilde{\mathbf{x}})$] will be proportional not only in the realised sample but also in repeated samples. Note that this repeated sampling is conditional on $\mathbf{X} = \tilde{\mathbf{x}}$ but not on $\mathbf{M} = \tilde{\mathbf{m}}$.

Second, consider repeated sampling conditional on $\mathbf{X} = \tilde{\mathbf{x}}$ and $\mathbf{M} = \tilde{\mathbf{m}}$. Assume that either $\tilde{\mathbf{x}}$ is observed or the distribution of $t\{o(\mathbf{Y}, \tilde{\mathbf{m}}), \tilde{\mathbf{m}}\}$, given $\mathbf{M} = \tilde{\mathbf{m}}$ and $\mathbf{X} = \tilde{\mathbf{x}}$ implied by the distribution $\int f_\theta(\mathbf{y} \mid \mathbf{x} = \tilde{\mathbf{x}}) r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\, d\mathbf{y}$, does not depend on the value of the missing part of $\tilde{\mathbf{x}}$. In equations (2)–(5) and (10), $f_\theta(\mathbf{y})$ should be replaced by $f_\theta(\mathbf{y} \mid \mathbf{x} = \tilde{\mathbf{x}})$, and $f_\theta(\mathbf{u})$ in equation (9) should be replaced by $f_\theta(\mathbf{u} \mid \mathbf{x} = \tilde{\mathbf{x}})$. Theorem 3 then continues to apply if "$f_{\boldsymbol{\theta}}(\mathbf{y}) g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) > 0$" is replaced by "$f_\theta(\mathbf{y} \mid \mathbf{x} = \tilde{\mathbf{x}}) g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) > 0$ and $b(\mathbf{y}) = \tilde{\mathbf{x}}$". Moreover, the realised MCAR condition in Theorem 3 can be replaced by the following weaker condition: $\forall \boldsymbol{\phi}$, $g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) = g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}^*)$ $\forall \mathbf{y}, \mathbf{y}^*$ such that $b(\mathbf{y}) = b(\mathbf{y}^*) = \tilde{\mathbf{x}}$. In the special case of repeated-measures data with fully observed covariates and $\mathbf{X}$ being these covariates, the everywhere version of this weaker condition has been called "covariate-dependent MCAR" [22, 41].

7. DISCUSSION

In this article we have highlighted inconsistencies in the use of the terms "missing at random" and "likelihood inference", and clarified the conditions required for ignorability of the missingness mechanism. We urge those writing about missing data to be clearer with respect to the assumptions being used and to employ clear terminology when describing approaches to inference, in particular, to make the distinction between direct-likelihood and frequentist likelihood concepts.

Rubin [33] used the term "ignorable" to mean that two likelihoods, one derived from a model for the data alone and one derived from a joint model for the data and the missingness pattern, are proportional or that two sampling distributions, the "potentially incorrect" and correct conditional distributions, are equal. In Section 5 we explained how this implies that certain inferences for $\boldsymbol{\theta}$ from the two models are the same. In this interpretation, ignorability is a property of the *assumed* missingness model. Whether this assumed model is correctly specified is not relevant. This interpretation of "ignorability" may not be universal, however. As we saw in Section 3, some writers have omitted the parameter $\boldsymbol{\phi}$ from their definition of MAR. Rather than refer to a model for the missingness mechanism, they appear to have been referring to the "true" missingness mechanism (which is usually unknown). Such writers may have interpreted ignorability to mean that using $L_2(\boldsymbol{\theta})$ for frequentist likelihood (or frequentist Bayesian) inference will be valid, that is, will yield consistent MLEs (or posterior modes), consistent variance estimators, confidence (or credible) intervals with asymptotic nominal coverage, etc. Theorem 1 implies the following result. Suppose that the "true" missingness mechanism is $P(\mathbf{M} = \mathbf{m} \mid \mathbf{Y} = \mathbf{y})$ and that $P(\mathbf{M} = \mathbf{m} \mid \mathbf{Y} = \mathbf{y}) = P(\mathbf{M} = \mathbf{m} \mid \mathbf{Y} = \mathbf{y}^*)$ $\forall \mathbf{m}, \mathbf{y}, \mathbf{y}^*$ such that $o(\mathbf{y}, \mathbf{m}) = o(\mathbf{y}^*, \mathbf{m})$. A hypothetical analyst who knew this "true" missingness mechanism and wanted to make inference for $\boldsymbol{\theta}$ taking missingness into account would use the likelihood $\int f_\theta(\mathbf{y}) P(\mathbf{M} = \tilde{\mathbf{m}} \mid \mathbf{Y} = \mathbf{y}) r(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\mathbf{m}})\, d\mathbf{y}$ and, by so doing, obtain valid frequentist likelihood (or frequentist Bayesian) inference. Theorem 1 implies that $L_2(\boldsymbol{\theta})$ is proportional to this likelihood, and hence that valid frequentist likelihood (or frequentist Bayesian) inference would also be obtained using L_2.

Despite MAR plus distinctness of parameters being presented in Little and Rubin [24] as the definition of ignorability (Definition 6.4), Theorems 1 and 2 only give *sufficient* conditions for when it is appropriate to ignore the missingness mechanism when making direct-likelihood and Bayesian inferences, respectively. In the case of direct-likelihood inference, Theorem 1 is concerned with sufficient conditions for $L_{3,\phi}(\boldsymbol{\theta})$, the fixed-$\boldsymbol{\phi}$ likelihood for $\boldsymbol{\theta}$, to be proportional to $L_2(\boldsymbol{\theta})$, the likelihood for $\boldsymbol{\theta}$ ignoring the missing data mechanism. It is conceivable that, even in the absence of realised MAR, there may be a restricted set of $\boldsymbol{\phi}$ values for which $L_2(\boldsymbol{\theta})$ is proportional to $L_{3,\phi}(\boldsymbol{\theta})$, and for this restricted set to contain the "true" $\boldsymbol{\phi}$ value. If so, it would be appropriate to ignore the missingness mechanism even though realised MAR does not hold. Lu and Copas [25] showed that, when $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct *and* the family of distributions $f_\theta(\mathbf{y})$ form a complete class, everywhere MAR is necessary and sufficient for ignorability in frequentist likelihood inference. It is straightforward to adapt their proof to show that when $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct and the family of distributions $f_\theta(\mathbf{y} \mid o(\mathbf{y}, \mathbf{m}) = o(\tilde{\mathbf{y}}, \tilde{\mathbf{m}}))$ form a complete class, then realised MAR is necessary and sufficient for ignorability in direct likelihood inference (we include a proof in the Appendix). Furthermore, there may conceivably be other ways, apart from that of using a fixed-$\boldsymbol{\phi}$ likelihood, to extract a likelihood for $\boldsymbol{\theta}$ from L_1, ways which may not require realised MAR and parameter distinctness in order for the extracted likelihood to be proportional to L_2. In the case of Theorem 2, it is conceivable that independence of the posterior distributions for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ may be a stronger condition than is necessary, and it seems to still be an open question whether there are substantially weaker conditions under which it is appropriate to ignore the missingness mechanism when performing Bayesian inference.

Note that the concept of missing data has been generalised to that of "coarsened" data [10]. When data are coarsened, data values are not necessarily either observed or missing, instead one observes a set of values that is known to contain the realised values. Censored survival data are an example of coarsened data: a survival time may be known to be greater than a given (censoring) time but not known exactly.

We conclude with some brief remarks on the potential practical implications of this work. Our review of the literature on the theory of missing data methods has highlighted a number of inconsistencies and a lack of clarity with respect to key definitions such as MAR and ignorability. We believe that these issues have clouded the development and broader understanding of methods in this area, partly because they intersect in considerable measure with issues in the foundations of statistical inference. Although the original definition of MAR (our "realised MAR") provides a clear basis for thinking about direct likelihood and Bayesian inferences, the majority of statistical practice is concerned with frequentist evaluations. Even those who emphasise the Bayesian interpretation of particular analyses are generally interested in repeated-sampling performance of procedures. Incomplete data methods that do not explicitly model the missing data mechanism (i.e., that assume ignorability) cannot be guaranteed to perform validly in repeated samples except under an "everywhere" MAR assumption. The restrictiveness of this assumption does not seem to be well understood, especially in complex problems with nonmonotone patterns of missingness [29, 31]. More importantly, further work is needed on methods to more effectively and systematically characterise the potential sensitivity of inferences to departures from the MAR assumption. Meanwhile, users of missing data methods need to be reminded that methods that assume ignorability provide tractable analyses only at the cost of untestable assumptions.

It is also important to consider that when there are missing data, there is more than one possible target of inference. Diggle et al. [4] discuss alternative possible study objectives and targets of inference that are relevant to those objectives.

Much recent research in methods for handling missing data has considered issues that are specific to the structure of the problem. For example, missingness in outcomes poses different challenges than does missingness in covariate values, and longitudinal (repeated measures) data present specific issues of their own. We believe that it should be possible to tackle these problems with greater clarity if the fundamental assumptions about missing data mechanisms and their connection with the concept of ignorability are better understood.

APPENDIX

Here we show that when $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct and the family of distributions $f_\theta(\mathbf{y} \mid o(\mathbf{y},\mathbf{m}) = o(\tilde{\mathbf{y}},\tilde{\mathbf{m}}))$ form a complete class, then realised MAR is necessary and sufficient for ignorability.

Let $\bar{o}(\mathbf{Y},\mathbf{M})$ denote the subvector of $\mathbf{Y}$ consisting of the elements whose corresponding elements of $\mathbf{M}$ equal zero. So, $\bar{o}(\mathbf{Y},\mathbf{M})$ contains the missing elements of $\mathbf{Y}$. For any fixed value $\mathbf{m}$ of $\mathbf{M}$, $f_\theta(\mathbf{y})$ can be written as

$$(11)\quad f_\theta(\mathbf{y}) = f_{1,\theta}\{o(\mathbf{y},\mathbf{m})\} f_{2,\theta}\{\bar{o}(\mathbf{y},\mathbf{m}) \mid o(\mathbf{y},\mathbf{m})\}.$$

Thus, choosing $\mathbf{m} = \tilde{\mathbf{m}}$ in equation (11), L_1 can be written as

$$\begin{aligned} L_1(\boldsymbol{\theta},\boldsymbol{\phi}) &= \int f_{1,\theta}\{o(\mathbf{y},\tilde{\mathbf{m}})\} f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\mathbf{y},\tilde{\mathbf{m}})\} \\ &\qquad \cdot g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y} \\ &= f_{1,\theta}\{o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} \int f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} \\ &\qquad \cdot g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y} \end{aligned}$$

and L_2 can be written as $L_2(\boldsymbol{\theta}) = f_{1,\theta}\{o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\}$.

THEOREM. *Suppose that* $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$, *that* $f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\}$ *is complete, and that* $g_\phi(\tilde{\mathbf{m}} \mid \tilde{\mathbf{y}}) > 0$ *for all* $\boldsymbol{\phi} \in \Omega_\phi$. *Then* $L_1(\boldsymbol{\theta},\boldsymbol{\phi})$ *is proportional to* $L_2(\boldsymbol{\theta})$ *for any* $\boldsymbol{\phi} \in \Omega_\phi$ *if and only if realised MAR holds.*

PROOF. The "if" argument holds because of Theorem 1. So, consider the "only if" argument. Suppose that $L_1(\boldsymbol{\theta},\boldsymbol{\phi})$ is proportional to $L_2(\boldsymbol{\theta})$ for any $\boldsymbol{\phi} \in \Omega_\phi$. Then it must be true that for all $\boldsymbol{\phi} \in \Omega_\phi$,

$$(12)\quad \int f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y}$$

cannot depend on $\boldsymbol{\theta}$. Hence, we can denote expression (12) as $Q\{\tilde{\mathbf{m}}, o(\tilde{\mathbf{y}},\tilde{\mathbf{m}}), \boldsymbol{\phi}\}$.

By definition,

$$\begin{aligned} &\int f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y} \\ &\quad - Q\{\tilde{\mathbf{m}}, o(\tilde{\mathbf{y}},\tilde{\mathbf{m}}), \boldsymbol{\phi}\} = 0. \end{aligned}$$

So,

$$\begin{aligned} &\int f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y} \\ &\quad - Q\{\tilde{\mathbf{m}}, o(\tilde{\mathbf{y}},\tilde{\mathbf{m}}), \boldsymbol{\phi}\} \\ &\qquad \cdot \int f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y} = 0 \end{aligned}$$

for all $\boldsymbol{\phi} \in \Omega_\phi$. It then follows that

$$\begin{aligned} &\int f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\} \\ &\quad \cdot [g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y}) - Q\{\tilde{\mathbf{m}}, o(\tilde{\mathbf{y}},\tilde{\mathbf{m}}), \boldsymbol{\phi}\}] r(\mathbf{y},\tilde{\mathbf{y}},\tilde{\mathbf{m}})\, d\mathbf{y} = 0 \end{aligned}$$

for all $\boldsymbol{\phi} \in \Omega_\phi$. So, if $f_{2,\theta}\{\bar{o}(\mathbf{y},\tilde{\mathbf{m}}) \mid o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})\}$ is complete, then $Q\{\tilde{\mathbf{m}}, o(\tilde{\mathbf{y}},\tilde{\mathbf{m}}), \boldsymbol{\phi}\} = g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y})$ for all $\boldsymbol{\phi} \in \Omega$ and for all $\mathbf{y}$ such that $o(\mathbf{y},\tilde{\mathbf{m}}) = o(\tilde{\mathbf{y}},\tilde{\mathbf{m}})$. Therefore, $g_\phi(\tilde{\mathbf{m}} \mid \mathbf{y})$ cannot depend on $\bar{o}(\tilde{\mathbf{y}},\tilde{\mathbf{m}})$, that is, the data are realised MAR. □

ACKNOWLEDGEMENTS

We thank Professor Mike Kenward for very useful discussions and for comments on a draft manuscript, and anonymous reviewers for their helpful suggestions.

S. R. Seaman is funded by MRC Grants U1052 60558 and MC_US_A030_0015. D. Jackson is funded by U1052 60558. J. B. Carlin and J. C. Galati wish to acknowledge the support of research Grant 607400 from the Australian National Health and Medical Research Council, and support provided to the MCRI from the Victorian Government's Operational Infrastructure Support Program.

REFERENCES

[1] ANSCOMBE, F. J. (1964). Normal likelihood functions. *Ann. Inst. Statist. Math.* **16** 1–19. MR0171348

[2] CLAYTON, D. and HILLS, M. (1993). *Statistical Models in Epidemiology*. Oxford Univ. Press, Oxford.

[3] COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

[4] DIGGLE, P., FAREWELL, D. and HENDERSON, R. (2007). Analysis of longitudinal data with drop-out: Objectives, assumptions and a proposal. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **56** 499–550. MR2405418

[5] DIGGLE, P. J. (2004). Estimation with missing data (correspondence). *Biometrics* **50** 580.

[6] EDWARDS, A. W. F. (1970). Discussion of "Application of likelihood methods to models involving large numbers of parameters" by J. D. Kalbfleisch and D. A. Sprott. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **32** 196–198.

[7] FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

[8] FITZMAURICE, G. M., LAIRD, N. M. and WARE, J. H. (2011). *Applied Longitudinal Analysis*, 2nd ed. Wiley, Hoboken, NJ. MR2830137

[9] HAREL, O. and SCHAFER, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96** 37–50. MR2482133

[10] HEITJAN, D. F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics* **49** 1099–1109.

[11] HEITJAN, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81** 701–708. MR1326420

[12] HEITJAN, D. F. (1997). Ignorability, sufficiency and ancillarity. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 375–381. MR1440587

[13] HEITJAN, D. F. (2004). Estimation with missing data (correspondence). *Biometrics* **50** 580.

[14] HEITJAN, D. F. and BASU, S. (1996). Distinguishing "missing at random" and "missing completely at random". *Amer. Statist.* **50** 207–213. MR1422070

[15] HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174

[16] HINDE, J. and AITKIN, M. (1987). Canonical likelihoods: A new likelihood treatment of nuisance parameters. *Biometrika* **74** 45–58. MR0885918

[17] JAEGER, M. (2005). Ignorability in statistical and probabilistic inference. *J. Artificial Intelligence Res.* **24** 889–917 (electronic). MR2200528

[18] KALBFLEISCH, J. D. and SPROTT, D. A. (1970). Discussion of "Application of likelihood methods to models involving large numbers of parameters". *J. R. Stat. Soc. Ser. B Stat. Methodol.* **32** 204–208.

[19] KASS, K. E. and WASSERMAN, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91** 1343–1370.

[20] KENWARD, M. G. and MOLENBERGHS, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statist. Sci.* **13** 236–247. MR1665713

[21] LITTLE, R. J. A. (1976). Comments on "Inference and missing data". *Biometrika* **63** 590–591.

[22] LITTLE, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90** 1112–1121. MR1354029

[23] LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York. MR0890519

[24] LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

[25] LU, G. and COPAS, J. B. (2004). Missing at random, likelihood ignorability and model completeness. *Ann. Statist.* **32** 754–765. MR2060176

[26] MOLENBERGHS, G. and KENWARD, M. G. (2007). *Missing Data in Clinical Studies*. Wiley, Chichester.

[27] MOLENBERGHS, G., KENWARD, M. G., VERBEKE, G. and BIRHANU, T. (2011). Pseudo-likelihood estimation for incomplete data. *Statist. Sinica* **21** 187–206. MR2796859

[28] PAWITAN, Y. (2001). *In All Likelihood*. Clarendon, Oxford.

[29] POTTHOFF, R. F., TUDOR, G. E., PIEPER, K. S. and HASSELBLAD, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Stat. Methods Med. Res.* **15** 213–234. MR2227446

[30] REID, N. (2000). Likelihood. *J. Amer. Statist. Assoc.* **95** 1335–1340. MR1825289

[31] ROBINS, J. M. and GILL, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Stat. Med.* **16** 39–56.

[32] ROYALL, R. M. (1997). *Statistical Evidence*: *A Likelihood Paradigm. Monographs on Statistics and Applied Probability* **71**. Chapman & Hall, London. MR1629481

[33] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

[34] RUBIN, D. B. (1976). Reply to comments on "Inference and missing data". *Biometrika* **63** 591–592.

[35] RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

[36] RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

[37] SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data. Monographs on Statistics and Applied Probability* **72**. Chapman & Hall, London. MR1692799

[38] STERNE, J. A., WHITE, I. R., CARLIN, J. B., SPRATT, M., ROYSTON, P., KENWARD, M. G., WOOD, A. M. and CARPENTER, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *Br. Med. J.* **338** art. no. b2393.

[39] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926

[40] TSOU, T.-S. and ROYALL, R. M. (1995). Robust likelihoods. *J. Amer. Statist. Assoc.* **90** 316–320. MR1325138

[41] WOOD, A. M., WHITE, I. R., HILLSDON, M. and CARPENTER, J. (2004). Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. *Int. J. Epidemiol.* **34** 89–99.