

Annual Review of Statistics and Its Application

Missing Data Assumptions

Roderick J. Little

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48105, USA;
email: rlittle@umich.edu

Annu. Rev. Stat. Appl. 2021. 8:89–107

First published as a Review in Advance on
August 21, 2020

The *Annual Review of Statistics and Its Application* is
online at statistics.annualreviews.org

<https://doi.org/10.1146/annurev-statistics-040720-031104>

Copyright © 2021 by Annual Reviews.
All rights reserved

**ANNUAL
REVIEWS CONNECT**

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

missing at random, ignorable missing data, Bayesian and frequentist inference, incomplete data, informative missingness, likelihood inference, missing-data mechanism, partially missing at random

Abstract

I review assumptions about the missing-data mechanisms that underlie methods for the statistical analysis of data with missing values. I describe Rubin's original definition of missing at random (MAR), its motivation and criticisms, and his sufficient conditions for ignoring the missingness mechanism for likelihood-based, Bayesian, and frequentist inference. Related definitions, including missing completely at random, always MAR, always missing completely at random, and partially MAR, are also covered. I present a formal argument for weakening Rubin's sufficient conditions for frequentist maximum likelihood inference with precision based on the observed information. Some simple examples of MAR are described, together with an example where the missingness mechanism can be ignored even though MAR does not hold. Alternative approaches to statistical inference based on the likelihood function are reviewed, along with non-likelihood frequentist approaches, including weighted generalized estimating equations. Connections with the causal inference literature are also discussed. Finally, alternatives to Rubin's MAR definition are discussed, including informative missingness, informative censoring, and coarsening at random. The intent is to provide a relatively nontechnical discussion, although some of the underlying issues are challenging and touch on fundamental questions of statistical inference.

1. INTRODUCTION: RUBIN'S DEFINITION OF MISSING AT RANDOM

In 1974 I was a statistics doctoral student at Imperial College, London, and Professor D.R. Cox, the editor of *Biometrika*, handed me my first journal article to referee. It was called “Inference and Missing Data” and was written by a statistician called Don Rubin, whom I had never heard of—I knew very little about the field of statistics at that time. Rubin’s paper treated the missingness indicator variables as part of a statistical model and laid out sufficient conditions for ignoring the missingness mechanism for different types of statistical inference. I struggled mightily with the paper, in part because of some of the subtle questions that I review here. Despite the reservations in my lengthy referee’s report, Professor Cox wisely decided to publish the paper, which became a highly cited, landmark article on statistical inference with missing data (Rubin 1976), together with a discussion from the referee (Little 1976). Rubin and I later collaborated on our book on missing data, now in the third edition (Little & Rubin 2019). In that edition, we updated our discussion of assumptions about missingness in previous editions, and the topic continues to create some confusion and controversy. The main goal of this article is to describe and hopefully shed light on these assumptions.

Before Rubin’s seminal paper, missing data were generally handled by simple ad hoc methods like complete-case (CC) analysis, where units with missing values are deleted and analysis is based on the remaining cases, or naïve imputation, where missing values of a variable are replaced by estimates such as the means of recorded values of that variable, predictions from a regression on observed variables, or the last recorded value in longitudinal studies. Early systematic attempts to account for missing data in statistical inferences were based on the method of maximum likelihood (ML) (see, for example, Anderson 1957, Hartley 1958, Trawinski & Bargmann 1964, Afifi & Elashoff 1966). In Section 2, I discuss assumptions underlying likelihood-based methods for handling inference from data sets with missing values; these methods include ML, Bayesian methods, and model-based multiple imputation (MI), where multiple data sets are created with missing values replaced by draws from their predictive distribution, with inference based on Rubin’s (1987) MI combining rules.

In Section 3, I discuss assumptions underlying other inferential approaches—namely, CC analysis and weighted CC analysis where units are weighted by the inverse of estimates of their response probabilities. These methods include inverse-probability weighted generalized estimating equations (GEE) and extensions where the estimates are augmented by predictions of the missing values.

In Section 4, I discuss some alternative definitions of missingness mechanisms, including informative missing data, definitions of MAR for parameter subsets, and ignorable coarsening mechanisms. Section 5 provides some concluding remarks.

2. LIKELIHOOD INFERENCE WITH MISSING DATA

2.1. Introduction

Let $Y = (y_{ij})$ denote an $(n \times p)$ rectangular data set without missing values, with i th row $y_i = (y_{i1}, \dots, y_{ip})$ where y_{ij} is the value of variable Y_j for unit i . With no missing values, likelihood inference is based on a statistical model for Y , which asserts that Y is sampled from a distribution with density $f_Y(y | \theta)$ indexed by unknown parameters θ . The notation omits any fixed covariates X for simplicity. The likelihood of θ is

$$L_Y(\theta | \tilde{y}) = \text{const.} \times f_Y(\tilde{y} | \theta),$$

where, here and elsewhere, a tilde refers to the realized value of a random variable, so that \tilde{y} denotes the observed value of Y . The argument of the likelihood L_Y is θ rather than \tilde{y} , and the constant can

depend on \tilde{y} but not on θ . The ML estimate of θ maximizes this function. Large-sample inference can be based on a normal distribution centered at the ML estimate, with a covariance matrix based on the observed or expected information matrix, or some approximation thereof. Bayesian inference is based on the posterior distribution of θ , which is proportional to $L_Y(\theta|\tilde{y})$ multiplied by a prior distribution $\pi(\theta)$ for the parameters.

With missing data, define the response indicator matrix $R = (r_{ij})$, such that $r_{ij} = 1$ if y_{ij} is observed and $r_{ij} = 0$ if y_{ij} is missing; alternatively, $1 - r_{ij}$ is called the missingness indicator for y_{ij} . Here, an unobserved value of y_{ij} is considered missing if it would be meaningful for analysis if observed (Little & Rubin 2019), in the sense of being relevant for the question of interest. For example, in a repeated measures setting, blood pressure data that are not observed because participants die are not considered missing since they are not meaningful for analysis—literally, their blood pressures are zero, but including values of zero is clearly not useful, and imputing a “blood pressure if the individual had not died” is at best questionable. Fully observed covariates X can be fixed and are implicit in the notation, but covariates with missing values need to be included in the set of Y variables.

Using the notation¹ in Mealli & Rubin (2015), let $Y_{(1)}$ denote the observed components of Y and let $Y_{(0)}$ denote the missing components of Y . In a slightly informal notation, rewrite $f_Y(y|\theta)$ as $f_Y(y_{(1)}, y_{(0)}|\theta)$. The natural likelihood with missing data is obtained by integrating the missing values $y_{(0)}$ of Y out of the density $f_Y(y_{(1)}, y_{(0)}|\theta)$ —that is,

$$L_{\text{ign}}(\theta|\tilde{y}_{(1)}) = \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta) dy_{(0)}. \quad 1.$$

ML estimates then maximize this function with respect to θ . Estimates of precision and asymptotic normal-based tests and confidence intervals can be based on standard ML theory, applied to this likelihood. Unlike earlier ad hoc imputation approaches, these inferences take into account the loss of information from the missing data.

Rubin (1976) noted that this approach not only assumes that the model $f_Y(y|\theta)$ is well specified but also requires assumptions about the missingness mechanisms that led to missing values. Rubin (1976) described inferences based on the likelihood in Equation 1 as ignoring the missingness mechanism—hence the label “ign” for L in this equation. He articulated the assumptions implied by basing inference on Equation 1 by formulating a model for the missing-data mechanism, characterized by the (discrete) conditional distribution of R given Y , say $f_{R|Y}(R|Y, \phi)$, where ϕ denotes unknown parameters. The full likelihood based on the observed values $Y_{(1)} = \tilde{y}_{(1)}$, $R = \tilde{r}$ is then obtained by integrating the missing data $y_{(0)}$ out of the density of the joint distribution of Y and R —that is,

$$L_{\text{full}}(\theta, \phi|\tilde{y}_{(1)}, \tilde{r}) = \int f_Y(\tilde{y}_{(1)}, y_{(0)}|\theta) f_{R|Y}(\tilde{r}|\tilde{y}_{(1)}, y_{(0)}, \phi) dy_{(0)}, \quad 2.$$

considered as a function of the parameters (θ, ϕ) . Rubin called the missingness mechanism ignorable for likelihood inference if inference about θ based on Equation 1 was the same as inference about θ based on Equation 2.

Rubin then specified sufficient conditions under which various forms of inference can be based on the ignorable likelihood (Equation 1) rather than the full likelihood (Equation 2). I now discuss these conditions and assumptions.

¹The notation y_{obs} and y_{mis} for observed and missing data is evocative and has been used extensively by myself and others, but is avoided here since it can be misinterpreted as implying conditioning on the set of observed variables; see Seaman et al. (2013) or Little & Rubin (2019).

2.2. Sufficient Conditions for Ignoring the Missingness Mechanism for Bayesian Inference

Sufficient conditions for ignoring the missingness mechanism are easiest to understand for pure likelihood inference, meaning inference based on the likelihood function (Equation 1) and ratios of likelihoods for different values of the parameters. I focus here on Bayesian inference, the most widely applicable form of pure likelihood inference, while noting that non-Bayesian versions of pure likelihood inference do have advocates. The sufficient conditions for Bayesian inference are contained in the following lemma.

Lemma 1. The following two conditions are sufficient for ignoring the missingness mechanism for Bayesian inference:

1. The missing data are missing at random (MAR), defined as

$$\begin{aligned} f_{R|Y}(R = \tilde{r} | Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y_{(0)}, \phi) \\ = f_{R|Y}(R = \tilde{r} | Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y_{(0)}^*, \phi) \text{ for all } y_{(0)}, y_{(0)}^* \text{ and } \phi. \end{aligned} \quad 3.$$

In other words, the function $f_{R|Y}(R = \tilde{r} | Y_{(1)} = \tilde{y}_{(1)}, Y_{(0)} = y_{(0)}, \phi)$, considered as a function of $Y_{(1)}$ and $Y_{(0)}$, can depend on the observed data, $Y_{(1)} = \tilde{y}_{(1)}$, but not on the missing data, $Y_{(0)}$.

2. The parameters θ and ϕ are assumed to be a priori independent.

Proof. Suppose we adopt a full model for Y and R as in Equation 2 and further assume a prior distribution $\pi(\theta, \phi)$, where θ and ϕ are a priori independent—that is,

$$\pi(\theta, \phi) = \pi_1(\theta)\pi_2(\phi).$$

Under MAR, Equation 3, the full likelihood, factorizes as

$$L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{r}) = L_{\text{ign}}(\theta | \tilde{y}_{(1)}) \times L_{\text{rest}}(\phi | \tilde{y}_{(1)}, \tilde{r}) \text{ for all } \theta, \phi. \quad 4.$$

Under (a) and (b), the full posterior distribution is

$$\begin{aligned} p(\theta, \phi | \tilde{y}_{(1)}, \tilde{r}) &= \text{const.} \times \pi(\theta, \phi) \times L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{r}) \\ &= \text{const.} \times [\pi_1(\theta) \times L_{\text{ign}}(\theta | \tilde{y}_{(1)})] \times [\pi_2(\phi) \times L_{\text{rest}}(\phi | \tilde{y}_{(1)}, \tilde{r})] \text{ for all } \theta, \phi. \end{aligned} \quad 5.$$

Thus, θ and ϕ are a posteriori independent, and the posterior distribution of θ is

$$p(\theta | \tilde{y}_{(1)}, \tilde{r}) = \text{const.} \times \pi_1(\theta) \times L_{\text{ign}}(\theta | \tilde{y}_{(1)}) \text{ for all } \theta,$$

that is, this posterior distribution is based on the ignorable likelihood, which does not involve the model for the missingness mechanism. \square

In large samples and for prior distributions that do not restrict the parameter space, the posterior distribution of the parameters is dominated by the likelihood. The posterior distribution of θ ignoring the missingness mechanism can then be based on the normal approximation:

$$(\theta | \text{data}) \sim N\left(\hat{\theta}, I_{\text{ign}(\theta\theta)}^{-1}\right), \quad 6.$$

where $N(a, B)$ denotes the multivariate normal distribution with mean a and covariance matrix B , $\hat{\theta}$ maximizes the ignorable likelihood $L_{\text{ign}}(\theta | \tilde{y}_{(1)})$, and $I_{\text{ign}(\theta\theta)}$ is the observed information

$$I_{\text{ign}(\theta\theta)} = -D_{\theta\theta} (L_{\text{ign}}(\theta | \tilde{y}_{(1)})) \big|_{\theta=\hat{\theta}}, \quad 7.$$

where $D_{\theta\theta}$ denotes the second derivative of the argument with respect to θ . For large-sample Bayesian inference, the conditions of Lemma 1 for ignoring the missingness mechanism can be weakened:

Lemma 2. The following are sufficient conditions for large-sample Bayesian inference based on Equation 6, ignoring the missingness mechanism:

1. The missing data are MAR, as in Equation 3, and
- 2*. (Distinctness). The parameters θ and ϕ are distinct, in the sense of having distinct parameter spaces. Formally, if $\Omega_{(\theta,\phi)}$ is the parameter space of θ and ϕ , then $\Omega_{(\theta,\phi)} = \Omega_\theta \times \Omega_\phi$.

The condition (2*) is weaker than a priori independence (2) because it is necessary but not sufficient for θ and ϕ to be a priori independent—even if θ and ϕ have distinct parameter spaces, they could be assigned dependent prior distributions. If the distinctness condition seems abstract, in many situations it amounts to the parameters θ and ϕ not having a subset of parameters in common. For example, in shared parameter models for longitudinal data (e.g., Little 1995), the set of parameters that control missingness are also included as parameters in the model for the variables Y . It is often reasonable to model data so that distinctness holds; if MAR holds but Distinctness is violated, then Bayesian inference based on Equation 1 remains valid if interpreted as excluding the information about θ from the missingness mechanism. For these reasons, MAR is a more important condition than distinctness in practice.

I now consider asymptotic frequentist inference based on ML. Forms of small-sample frequentist inference are considered in Section 3.

2.3. Sufficient Conditions for Ignoring the Missingness Mechanism for Asymptotic Frequentist Maximum Likelihood Inference

Asymptotic frequentist ML inference ignoring the missingness mechanism can be based on the following approximation of the sampling distribution of $\hat{\theta}$:

$$(\hat{\theta} \mid \text{data}) \sim N\left(\theta, I_{\text{ign}(\theta\theta)}^{-1}\right), \quad 8.$$

where the roles of θ and $\hat{\theta}$ in Equation 6 are reversed. In Equation 8, the covariance matrix is based on the observed information, Equation 7; I consider other choices later. A difficult question is whether the MAR and distinctness conditions of Section 2.1 are also sufficient for frequentist inference based on Equation 8 to be valid.

The definition of MAR (Equation 3) is for all values of the unknown quantities $y_{(0)}$ and ϕ , for fixed $(\bar{y}_{(1)}, \bar{r})$. To quote Mealli & Rubin (2015, pp. 996–97), Equation 3 is “a statement about evaluating a function at a specific value of the indicator for observed and missing values, \bar{r} , and of the observed data values, $\bar{y}_{(1)}$, to see if it varies with possible values of the missing values, $y_{(0)}$, for any value of the parameter ϕ ; it is not a statement about conditional independence.” Unlike Equation 3, Equation 8 concerns the distribution of $\hat{\theta}$ in repeated sampling, with values of $(y_{(0)}, y_{(1)}, r)$ other than those in the realized data set. For the sampling distribution of $\hat{\theta}$ not to involve the model for the missingness mechanism, the following stronger condition, which Mealli & Rubin (2015) call missing always at random (MAAR), is required:

$$\begin{aligned} f_{R \mid Y}(R = r \mid Y_{(1)} = y_{(1)}, Y_{(0)} = y_{(0)}, \phi) \\ = f_{R \mid Y}(R = r \mid Y_{(1)} = y_{(1)}, Y_{(0)} = y_{(0)}^*, \phi) \text{ for all } r, y_{(1)}, y_{(0)}, y_{(0)}^* \text{ and } \phi. \end{aligned} \quad 9.$$

The “always” here refers to all possible values of $(y_{(0)}, y_{(1)}, r)$ in repeated sampling from the distribution of $\{R, Y\}$. Seaman et al. (2013) make a similar distinction between Equations 3 and 9, calling the condition in Equation 3 realized MAR and the condition in Equation 9 everywhere MAR. They discuss confusion in the literature arising from failure to distinguish between the two conditions, and they describe MAAR and distinctness as sufficient conditions for ignoring the missingness mechanism for asymptotic frequentist ML inference.

MAAR is a stronger condition than MAR—that is, MAAR implies MAR but MAR does not imply MAAR—and it is harder to assess because we don't know which patterns of missing values and associated values of $y_{(0)}$ might arise in repeated sampling. I argue in Lemma 3 that MAR remains sufficient for asymptotic frequentist ML inference, even though the sampling distribution of $\hat{\theta}$ might depend on the missingness mechanism in that case. Kenward & Molenberghs (1998) make a similar argument, informally, without proof.

Lemma 3. The conditions 1 and 2 of Lemma 1, namely, MAR and distinctness, are sufficient for ignoring the missingness mechanism for asymptotic frequentist ML inference with precision based on the inverse of the observed information, as in Equation 8.

Proof. Suppose we posit a full model for Y and R , with distinct and identified parameters (θ, ϕ) . Asymptotic inference could then be based on the normal approximation:

$$\begin{pmatrix} \theta - \hat{\theta} \\ \phi - \hat{\phi} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} I_{\text{full}(\theta\theta)} & I_{\text{full}(\theta\phi)} \\ I_{\text{full}(\phi\theta)} & I_{\text{full}(\phi\phi)} \end{pmatrix}^{-1} \right], \quad 10.$$

where $(\hat{\theta}, \hat{\phi})$ are ML estimates of (θ, ϕ) , and

$$I_{\text{full}(\theta\theta)} = -D_{\theta\theta} (\log L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{r})) \big|_{\theta=\hat{\theta}, \phi=\hat{\phi}}, I_{\text{full}(\theta\phi)} = -D_{\theta\phi} (\log L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{r})) \big|_{\theta=\hat{\theta}, \phi=\hat{\phi}},$$

and $I_{\text{full}(\phi\phi)} = -D_{\phi\phi} (\log L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{r})) \big|_{\theta=\hat{\theta}, \phi=\hat{\phi}}$

are the components of the observed information evaluated at ML estimates of the parameters.

Under MAR, L_{full} factors, as in Equation 4, into the product of a term for θ and a term for ϕ . Given that θ and ϕ are distinct, the ML estimate $\hat{\theta}$ maximizes $L_{\text{ign}}(\theta | \tilde{y}_{(1)})$, and

$$I_{\text{full}(\theta\theta)} = -D_{\theta\theta} (\log L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{r})) = -D_{\theta\theta} (\log L_{\text{ign}}(\theta | \tilde{y}_{(1)})) = I_{\text{ign}(\theta\theta)}; I_{\text{full}(\theta\phi)} = 0.$$

Thus, tests and confidence intervals for θ based on Equation 10 are the same as corresponding tests and confidence intervals based on Equation 8, when evaluated using the observed data. This is despite the fact that the sampling distribution of $(\hat{\theta}, \hat{\phi})$ is only ignorable under the stronger MAAR assumption in Equation 9; for other repeated samples where MAR is violated, the observed information matrix in Equation 9 would not be block diagonal and the full model is needed, but for the realized sample, the information matrix is block diagonal under MAR and distinctness. Indeed, ML inference ignoring the missingness mechanism is not affected by misspecifying the model for the missingness mechanism, because this model does not enter into the inference for the data at hand. \square

Lemma 3 also holds with the covariance matrix in Equation 10 replaced by the inverse of the sandwich estimator, or the covariance matrix from bootstrap samples. However, the result would not hold if the observed information were replaced by the expected information, which in general requires the stronger assumption of missing completely at random (MCAR), discussed in Section 3. This is a serious limitation (Kenward & Molenberghs 1998).

2.4. Unit Missing at Random and Missing Always at Random

The definitions in Section 2 are for the matrices Y and R , but they are easier to interpret if we can assume independence between the individual units, or rows, of the matrix. Following Mealli & Rubin (2015), suppose now that the units are exchangeable and can be modeled as independent, conditional on parameters θ, ϕ and any fully observed variables X . That is,

$$f_{Y,R}(Y, R | X, \theta, \phi) = \prod_{i=1}^n f_Y(y_i | x_i, \theta) f_R(r_i | x_i, y_i, \phi),$$

where the product is over n units in the data matrix. For a general pattern of missingness, let \mathcal{P} denote the set of missing-data patterns that have positive probability of occurrence in repeated sampling, and let $\tilde{\mathcal{P}} \subset \mathcal{P}$ be the patterns in the actual data set. Then MAAR becomes unit MAAR,

$$f_{R|Y}(r_i | x_i, y_{(1i)}, y_{(0i)}, \phi) = f_{R|Y}(r_i | x_i, y_{(1i)}, y_{(0i)}^*, \phi) \text{ for all } r_i, x_i, y_{(1i)}, y_{(0i)}, y_{(0i)}^*, \phi \text{ and } i \in \mathcal{P}, \quad 11.$$

and MAR becomes unit MAR,

$$f_{R|Y}(r_i = \tilde{r}_i | \tilde{x}_i, \tilde{y}_{(1i)}, y_{(0i)}, \phi) = f_{R|Y}(r_i = \tilde{r}_i | \tilde{x}_i, \tilde{y}_{(1i)}, y_{(0i)}^*, \phi) \text{ for all } y_{(0i)}, y_{(0i)}^*, \phi \text{ and } i \in \tilde{\mathcal{P}}, \quad 12.$$

which is weaker than MAAR in that it is a condition on the function $f_{R|Y}$ only for the observed values of $(\tilde{r}_i, \tilde{y}_{(1i)}, \tilde{x}_i)$, $i = 1, \dots, n$.

The following example illustrates the distinction between unit MAR and MAAR in a well-known situation.

Example 1 (Comparing means of two independent normal samples with potentially missing data).

Consider the familiar example of comparing the means of two independent normal samples with the same variance. The hypothetical complete data consist of (y_i, x_i) , $i = 1, \dots, n$, where y_i is the outcome and $x_i = j$ for group j , $j = 1$ or 2 . The model for Y given X is that $(y_i | x_i = j, \theta) \sim_{\text{ind}} N(\mu_j, \sigma^2)$, the normal distribution with mean μ_j and variance σ^2 , and $\theta = (\mu_1, \mu_2, \sigma^2)$. With no missing data, classical frequentist inference (tests and confidence intervals) for the difference in means $\delta = \mu_2 - \mu_1$ is based on the pivotal quantity $t = ((\bar{y}_2 - \bar{y}_1) - \delta) / (s\sqrt{1/n_1 + 1/n_2})$, which has a Student's t -distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom; here \bar{y}_j is the sample mean of Y and n_j is the number of units with $x_i = j$, and s^2 is the pooled sample variance. In particular, a 95% confidence interval for δ is

$$I_{0.95}(\delta) = \bar{y}_2 - \bar{y}_1 \pm t_{\nu, 0.975} \left(s\sqrt{1/n_1 + 1/n_2} \right), \quad 13.$$

where $t_{\nu, 0.975}$ is the 97.5 percentile of the t -distribution with ν degrees of freedom. Equation 13 is also the 95% Bayesian credible interval under the Jeffreys prior distribution $\pi(\mu_1, \mu_2, \log \sigma^2) \propto \text{const}$.

Suppose X is always observed, but Y potentially has missing values, so $r_i = 1$ if y_i is observed and $r_i = 0$ if y_i is missing. We assume that

$$\Pr(r_i = 1 | x_i, y_i, \phi) = b_1(x_i, \phi), \quad 14.$$

where b_1 is an unknown function. This means that missingness is allowed to depend on X , that is, the response rates in the two groups can differ, but missingness does not depend on the value of Y . Then missingness is MAR and MAAR, since the probability of response depends on the group indicator, which is always observed. Because units with y_i missing provide no information for the regression of Y on X , the usual t -inference methods can be applied to the set of units where Y is observed, provided at least one unit is observed in each group and at least three units overall, so that the pooled variance has at least one degree of freedom.

Suppose now that we allow missingness to depend on both X and Y , as in:

$$\Pr(r_i = 1 | x_i, y_i, \phi) = b_1(x_i, y_i, \phi).$$

As a concrete example, suppose that missingness is a form of censoring on Y in the group with $X = 2$. That is,

$$b(x_i, y_i, \phi) = \begin{cases} 1, & \text{if } x_i = 1 \\ 1, & \text{if } x_i = 2 \text{ and } y_i \leq \phi, \\ 0, & \text{if } x_i = 2 \text{ and } y_i > \phi. \end{cases} \quad 15.$$

For example, the values of Y in group 2 are measured using a flawed instrument that does not return values greater than an unknown ϕ . This mechanism is not unit MAAR, but it is unit MAR if all the n values of y_i happen to be observed in the realized data set. In that case the data set has no missing values. The standard t -based confidence interval (Equation 13) for δ can be computed, but it is not valid in a frequentist sense since, in repeated samples, missing values of Y in groups with $X = 2$ are missing not at random (MNAR). The sample mean \bar{y}_2 is a biased estimate of μ_2 , with a bias that depends on ϕ ; as ϕ increases, the bias decreases, and the likelihood that no values of Y are missing increases. Asymptotically \bar{y}_2 is consistent for μ_2 and the standard frequentist analysis is valid, but a satisfactory small-sample frequentist approach seems lacking, particularly when ϕ is unknown.

Although not a valid 95% confidence interval, the standard t -interval (Equation 13) is a valid 95% credibility interval in a Bayesian sense because it is valid under MAR provided that θ and ϕ are a priori independent, often a reasonable assumption. With MNAR missing values, a Bayesian analysis can be constructed by adding a prior distribution for ϕ (see, for example, Little & Rubin 2019, chapter 15), although (as with any Bayesian analysis) its small sample validity is dependent on correctness of the model and judicious choices of prior distributions.

In summary, for the mechanism (Expression 15) but all the values of Y observed, the standard analysis is valid for pure likelihood or Bayesian inference, but the frequentist analysis is invalid because MAAR is violated. For a similar example, readers are directed to Heitjan (2004). This example reflects some of the general philosophical debates about the relevance of hypothetical samples that are not observed for the inference based on the sample at hand.

The literature lacks comparisons of MAR and MAAR for more general patterns. Mealli & Rubin (2015) suggested that for a general pattern, MAAR implies that missingness can only depend on fully observed variables, but in a correction, Mealli & Rubin (2016) noted that this assertion required the (strong) assumption of conditional independence of the response indicators for each variable given data Y .

2.5. The Nature of Unit Missing at Random and Its Meaning for Monotone and Nonmonotone Missingness Patterns

In this section I discuss the meaning of unit MAR for some simple patterns of missing data and make some additional comments.

Example 2 (MAR for monotone data). Suppose the data on p variables (Y_1, \dots, Y_p) can be organized (possibly after reordering the variables) to have the monotone data pattern of **Figure 1**. In words, if the k th variable is missing for unit i , that is, $\tilde{r}_{ik} = 0$, then $\tilde{r}_{i(k+1)} = \dots = \tilde{r}_{ip} = 0$, that is, the variables $k+1, \dots, p$ are also missing. A special case is univariate nonresponse, where missingness is confined to the last variable. The columns in **Figure 1** could be multivariate. This monotone missingness pattern arises from attrition in longitudinal studies, where units drop out prior to the end of the study and do not return. Many longitudinal studies have a prevailing monotone pattern, except for a few cases that are missing intermittently.

It is easily seen that missingness for a monotone pattern is unit MAR if response to the k th variable depends on the variables $1, \dots, k-1$ (the history up to k) but not on variables $k, k+1, \dots, p$. That is,

$$\Pr(\tilde{r}_{ik} = 1 \mid \tilde{r}_{i1} = \tilde{r}_{i2} = \dots = \tilde{r}_{i,k-1} = 1, \tilde{y}_{i(1)}, y_{i(0)}, \phi) = b_k(\tilde{y}_{i1}, \dots, \tilde{y}_{i,k-1}, \phi), k = 2, \dots, p,$$

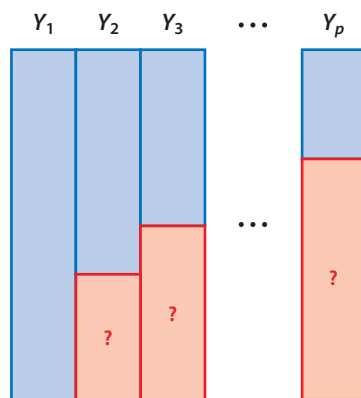


Figure 1

Monotone pattern of missing data.

for any functions b_2, \dots, b_p .

From a causal perspective, the idea that missingness at time t can depend only on history up to time t seems understandable. The term “sequentially ignorable nonresponse” is sometimes used to describe MAR for a monotone missingness pattern.

Example 3 (MAR for nonmonotone data: bivariate data with a general pattern). Unit MAR for nonmonotone data is less intuitive. Consider the simplest case of bivariate data with a general missingness pattern. Let $Y = [(y_{i1}, y_{i2}), i = 1, \dots, n]$ denote an independent sample from two variables Y_1, Y_2 with probability density $f_Y(y_{i1}, y_{i2} | \theta)$ indexed by unknown parameters θ . If all four patterns arise in the data, unit MAR implies that, for all y_i ,

$$\begin{aligned} \Pr(r_i = (0, 0) | \tilde{y}_{i(1)}, y_{i(0)}) &= b_{00} \\ \Pr(r_i = (1, 0) | \tilde{y}_{i(1)}, y_{i(0)}) &= b_{10}(\tilde{y}_{i1}) \\ \Pr(r_i = (0, 1) | \tilde{y}_{i(1)}, y_{i(0)}) &= b_{01}(\tilde{y}_{i2}) \\ \Pr(r_i = (1, 1) | \tilde{y}_{i(1)}, y_{i(0)}) &= 1 - b_{00} - b_{10}(\tilde{y}_{i1}) - b_{01}(\tilde{y}_{i2}). \end{aligned}$$

The last equation arises because the four probabilities sum to one (Little & Rubin 2019, example 1.19).

For a nonmonotone pattern like this, the causal interpretation of MAR is more fraught. In particular, consider this pattern of missingness when Y_1 and Y_2 are ordered sequentially over time. MAR implies that missingness of Y_2 can depend on the prior variable Y_1 for cases where Y_1 is observed, and missingness of Y_1 can depend on the future variable Y_2 for cases where Y_2 is observed. This mechanism seems difficult to reconcile causally.

Robins & Gill (1997) define randomized monotone missingness (RMM) models and argue that they represent the most general plausible physical mechanism for generating nonmonotone ignorable data. They argue that ignorable mechanisms that are not RMM are inappropriate since there is no physical mechanism for generating them. However, MAR for a nonmonotone pattern is a weaker assumption than, say, sequential ignorability, so it may approximate an actual MNAR mechanism more closely. Some have argued for omitting variables in an imputation model that are not exogenous in the substantive causal model of interest. I believe this is a mistake because the goal in imputation is to predict the missing values, with causal considerations entering in the analysis model for the imputed data.

There has been considerable debate about the interplay between causal inference and missing data—some arguing that everything is a causal inference problem and others saying that everything is a missing-data problem. On the one hand, the Neyman/Rubin causal model defines causal effects as a missing-data problem, where outcomes are defined under alternative treatments and outcomes under treatments not assigned are missing (see, e.g., Rubin 1978, Imbens & Rubin 2015). On the other hand, modeling the missingness mechanism is clearly a causal exercise, even if the final inference is descriptive rather than causal. Advocates of graphical causal models (e.g., Pearl et al. 2016) argue that these models enhance the ability to elucidate and determine restrictions that might identify MNAR models. The idea may have merit, although more real-world examples would help to make the case to this observer.

I conclude this section with some other general comments about MAR.

1. MAR is often loosely described as “missingness does not depend on the missing data $y_{(0)}$ after conditioning on the observed data $\tilde{y}_{(1)}$.” This formulation is a statement of conditional independence and, as such, describes MAAR rather than MAR, as discussed in Section 2.3. The formulation works best for simple missingness patterns, such as when missingness is confined to a single variable. In more complex cases, such as Example 3 above, it does not capture the nuances of MAR so well because the nature of the conditioning on observed data varies across patterns.
2. Some have objected to Rubin’s definition as unintuitive, particularly for nonmonotone missingness patterns like Example 3 above; however, Rubin based his definition of MAR not on intuitive notions, but rather as the weakest general condition under which pure likelihood inference can be based on $L_{\text{ign}}(\theta \mid \tilde{y}_{(0)})$ without needing a model for the missingness mechanism. An advantage of this approach is that intuitive notions of MAR may differ. I think attempts to redefine MAR are likely to compound rather than alleviate confusion.
3. The unit MAR definition may appear to conflict with the intuitive idea that missingness may be random for some units, but not random for others. For example, in a household survey of individuals, some units in the sample may be unit nonrespondents for reasons unconnected with survey variables and hence MAR—for example, they did not answer the door because their alarm clock failed to wake them. Others, in contrast, may not respond for reasons related to the survey and hence are not MAR—for example, they failed to respond to a health survey because they were hospitalized and hence not at home. Expressions 11 and 12 should be interpreted as conditioning on the variables that are observed for some units. When information on reasons for missingness is available, it is a good idea to include it via covariates in the analysis. Latent reasons for missingness are not conditioned on, so the different cases mentioned above are pooled, resulting in aggregate response probabilities. Latent ignorable models, where missingness is modeled as a function of latent variables, have been proposed in the literature (e.g., Barnard et al. 2002, Peng et al. 2004, Harel & Schafer 2009, Beesley et al. 2019) and can be useful in particular situations. Since values of the latent variable are not measured, these are examples of MNAR models.
4. MAR and distinctness are sufficient but not always necessary for ignoring the missingness mechanism for pure likelihood inference—Rubin (1976) describes them as the weakest conditions that can be applied generally, but in specific settings they can be weakened. The following example is in Little et al. (2016).

Example 4 (An example where data are MNAR but the missingness mechanism is ignorable). Suppose a random sample of size n is drawn from a finite population of size N . Two variables, Y_1 and Y_2 , are both observed ($r_i = 1$) for $i = 1, \dots, m$ and both missing ($r_i = 0$) for $m + 1, \dots, n$. Denote the resulting respondent data as Y_{resp} . Suppose units are

assumed to be independently distributed with density

$$f_Y(y_{i1}, y_{i2} | \theta) = f_1(y_{i1} | \theta_1) \times f_2(y_{i2} | y_{i1}, \theta_2),$$

where θ_1 and θ_2 are distinct parameters; also,

$$f_{R|Y}(r_i = 1 | y_{i1}, y_{i2}, \phi) = f_{R|Y}(r_i = 1 | y_{i1}, \phi) \text{ for all } y_{i1}, y_{i2}, \phi, \quad 16.$$

so that missingness of Y_1 and Y_2 depends on Y_1 but not Y_2 . Auxiliary data on the marginal distribution of Y_1 are also available for the whole population, say, $Y_{\text{aux}} = [y_{j1}^*, j = 1, \dots, N]$, so $y_{(1)} = [Y_{\text{resp}}, Y_{\text{aux}}]$. The vector Y_{aux} includes the respondent values of Y_1 in Y_{resp} , but the linkage between Y_{aux} and Y_{resp} is broken (e.g., De Groot & Goel 1980), in the sense that the set of m indices $i = 1, \dots, m$ in Y_{resp} are an unknown set of the population indices $j = 1, \dots, N$. Data of this form arise commonly in sample surveys, where the auxiliary data are available for the entire population from a census.

The missingness mechanism (Equation 16) is MNAR because missingness of unit i depends on the value y_{i1} , which is missing for the incomplete cases. However, Little et al. (2016) show that the missingness mechanism is ignorable because the parameters θ_1 governing the marginal distribution of Y_1 can be estimated from Y_{aux} , and the parameters θ_2 governing the conditional distribution of Y_2 given Y_1 can be estimated from Y_{resp} , without modeling the missingness mechanism.

2.6. Approaches to Likelihood-Based Inference with Missing Data

If we assume that missing data are MAR, analyses can be based on the ignorable likelihood, avoiding the need to model the missingness mechanism. The main approaches are (a) asymptotic inference based on the method of ML; (b) Bayesian inference based on the posterior distribution obtained by multiplying the likelihood by a prior distribution for the model parameters; and (c) MI, a variant of Bayesian inference where missing values are replaced by draws from their predictive distribution, and inferences are based on Rubin's MI combining rules (Rubin 1987). Nearly all the software for analysis of data with missing values does not model the missingness mechanism and hence implicitly assumes MAR. Little & Rubin (2019) provide extensive examples.

Two key and related challenges are that the MAR assumption cannot be directly tested from the observed data without additional structural assumptions, and that MNAR models are often difficult to specify correctly and may involve parameters that are unidentified or poorly identified from the data. Little & Rubin (2019, chapter 15) describe five approaches that address the problems with fitting MNAR models:

1. Follow up a sample of nonrespondents and incorporate this information into the main analysis.
2. Adopt a Bayesian approach, assigning the parameters prior distributions. Bayesian inference does not generally require that the parameters are identified, although inferences may be weak and sensitive to the choice of prior distribution.
3. Impose restrictions on model parameters, specifically, setting to zero coefficients in the models for the outcome or response.
4. Conduct analysis to assess sensitivity of inferences for quantities of interest to different choices of the values of parameters poorly estimated from the data.
5. Selectively discard data to avoid modeling the missingness mechanism. Sometimes a MNAR missingness mechanism can be ignorable after discarding some observations (see Little & Zhang 2011, Little et al. 2016).

3. SOME NON-LIKELIHOOD-BASED APPROACHES TO THE ANALYSIS OF DATA WITH MISSING VALUES

3.1. Introduction

Seaman et al. (2013) distinguish between (a) frequentist ML inference based on the asymptotic distribution of the ML estimate $\hat{\theta}$ of θ and (b) other forms of frequentist inference. I discuss sufficient conditions for *a* in Section 2.3. Sufficient conditions for *b* are harder to elucidate since it is less clear what it means—any form of inference, including Bayes, can be viewed with a frequentist lens, in the sense that it has frequentist properties in repeated sampling. Rubin (1976) showed² that for general frequentist inference to be valid it was sufficient that the data are MCAR, defined as

$$f_{R|Y}(R = \tilde{r} | y, \phi) = f_{R|Y}(R = \tilde{r} | y^*, \phi) \text{ for all } y, y^* \text{ and } \phi. \quad 17.$$

Note that, as with the definition of MAR, MCAR does not assume independence of R and Y because the equality in Equation 17 is only for the observed value \tilde{r} of R . Independence is implied by the condition called missing always completely at random (MACAR) by Mealli & Rubin (2015):

$$f_{R|Y}(R = r | y, \phi) = f_{R|Y}(R = r | y^*, \phi) \text{ for all } r, y, y^* \text{ and } \phi. \quad 18.$$

To illustrate the difference between MACAR and MCAR, consider a sample survey with two types of nonresponse: unit nonresponse, where all the survey variables are missing, and item nonresponse, where at least one of the survey variables is recorded. Suppose that unit nonresponse depends on the values of survey variables, but item nonresponse is completely at random; that is, it does not depend on any of the survey variables. In the observed sample, there happens to be no unit nonresponse; that is, the only missing values are from item nonresponse. Then the missingness mechanism for that survey is MCAR, but it is not MACAR, because unit nonresponse occurs in hypothetical repetitions of the survey, and for those data sets the data are not MCAR.

Assuming independence of units, the MCAR condition becomes unit MCAR,

$$f_{R|Y}(R_i = \tilde{r}_i | y_i, \phi) = f_{R|Y}(R_i = \tilde{r}_i | y_i^*, \phi) \text{ for all } i, y_i, y_i^*, \text{ and } \phi.$$

There is some confusion in the literature about whether missingness that depends only on fixed covariates, as in Equation 14 in Example 1, is MCAR or MAR. It is MAR but not MCAR according to Rubin's (1976) original terminology, but some have defined it as MCAR. Arguments can be made on both sides of this terminological issue; I have opted to stick with Rubin's original framing and have used the term "covariate-dependent missingness" for this type of mechanism (Little 1995).

Most of the literature on incomplete data that is not based on likelihoods concerns weighted versions of GEE applied to the complete cases, and augmented versions that incorporate predictions of the missing values. I discuss this literature in the next section.

3.2. Complete-Case Analysis

A common strategy when faced with nonresponse is simply to discard units with missing values and analyze the remaining complete cases. This approach is justified if the information in the discarded incomplete cases for the estimand of interest is small relative to the information in the complete

²Rubin (1976) defined a condition, observed at random (OAR), which is no longer in common usage. MCAR is equivalent to the combination of both MAR and OAR; according to Mealli & Rubin (2015), MCAR was first defined in Marini et al. (1980).

cases. A common question is, How small is small enough? This is not straightforward, as it depends on the specifics of the analysis. Consider, for example, a situation where a variable Y has missing values, and a set of variables X is fully observed. The incomplete units with only X measured can have considerable information for inference about the marginal mean of Y , particularly if X is highly predictive of Y . However, if missingness depends only on the X s—that is, the missingness mechanism is MAR and the distinctness condition of Lemma 2 applies—then the incomplete units carry no information at all for the parameters of the regression of Y on X and can be discarded (Little & Rubin 2019, section 3.2).

The assumption of MCAR is often stated as being a requirement for the validity of CC analysis. It is sufficient for CC analysis to be valid, but whether it is necessary depends on the specifics of the analysis, and in particular on the estimand. In the example just discussed, CC analysis for the marginal distribution of Y generally requires MCAR, but CC analysis for the regression of Y on X remains valid under MAR; that is, it is valid if missingness depends on the covariates X . This continues to be true when there are missing values of the covariates, provided the probability of being complete does not depend on the outcome. That condition includes MNAR situations where missingness depends on values of covariates that are sometimes missing.

3.3. Weighted Complete-Case Analysis

In sample surveys where units have differing probabilities of selection, analyses generally weight units by their sampling weights, defined as the inverse of the probability of selection. With missing data, an analogous approach is nonresponse weighting, which weights the completely observed units by the inverse of their probability of response. Unlike probability sampling, these weights are unknown and need to be estimated from the data. A simple version of this is adjustment cell weighting, where cells are created with similar values of fully observed covariates, and the weights of respondents in a cell are proportional to the inverse of the response rate in that cell. More generally, weights can be the inverse of estimated response propensity from a regression of the response indicator on fully observed covariates. When applied to estimates from GEE, this strategy is called inverse-probability weighted GEE.

Weighted CC analysis is popular because it is relatively simple, but it has some important limitations:

1. It is based on an assumption of covariate missingness—that is, the probability of being complete is assumed to depend only on the fully observed covariates. For unit nonresponse, where values of the variables Y with missing values are either fully observed or completely missing, this is equivalent to MAR. For more general patterns of missing data, MAR is a weaker assumption than covariate missingness because it also allows missingness to depend on observed values of Y .
2. Unlike likelihood-based methods, weighted CC analysis is not efficient, and it can be highly inefficient if weights are highly variable and not predictive of the Y variable of interest.
3. Extensions of weighting to monotone missing data are possible (Little & David 1988), but there is no satisfactory generalization to nonmonotone patterns. The complete cases can still be weighted, but information in the incomplete cases is not exploited. This is one reason why weighting is not the preferred approach to item nonresponse in surveys, which generally does not have a monotone pattern.

Augmented inverse-probability weighted GEE is an extension of inverse-probability weighted GEE that creates predictions from a model to recover information in the incomplete units and applies inverse-probability weighted GEE to the residuals from the model. This approach achieves

a gain in efficiency under a strong prediction model and has the property known as double robustness, meaning that it yields consistent estimates if either the model used to create predictions of the missing values, or the model used to estimate the propensity to respond, is correct (Robins et al. 1995, Robins & Rotnitzky 1995, Lipsitz et al. 1999). The methods as implemented generally assume MAR, although weights can be created for MNAR models; see, in particular, Scharfstein et al. (1999) and the discussion of that paper.

Example 5 (Regression with an incomplete covariate). Suppose interest concerns the regression of Y on X_1, \dots, X_p , where one of the covariates, say X_1 , has missing values, and the other variables X_2, \dots, X_p, Y are fully observed. The incomplete cases with X_1 missing then have considerable information for the intercept and coefficients of X_2, \dots, X_p but very limited information for the coefficient of X_1 (Little 1992). The incomplete cases are thus of limited value if the primary interest is in the coefficient of X_1 , but they are of considerably more value if the primary interest is in other coefficients. In particular, if X_1 is weakly associated with Y , the incomplete cases have about as much information as the complete cases for these other regression coefficients.

If the missingness mechanism is MAR, and missingness depends on the outcome Y , then inverse-probability weighting of the complete cases can reduce bias in the estimated regression coefficients. However, if missingness depends on the covariates X but not on the outcomes, unweighted CC analysis yields consistent (though potentially inefficient) estimates of the regression coefficients, and inverse-probability weighting of the complete cases can be biased. Thus, the question of whether weighting leads to improved estimates of the regression coefficients depends on the missingness mechanism.

4. OTHER DEFINITIONS OF MISSINGNESS MECHANISMS

4.1. Informative Missingness

The related terms “informative missingness” and “informative censoring” and their relationship with the definitions of missingness mechanisms discussed above are a source of much confusion. Wu & Carroll (1988) introduced the term “informative censoring” in the course of modeling repeated-data measures data Y that are right censored or missing observations caused by the participant’s death or withdrawal. The idea is that the fact and time of death informs the relationship between Y and time T , modeled in the paper as linear with random individual slopes and intercepts. Note that this is not the same as censored data in the usual survival analysis setting, where survival time is right-censored for individuals still alive at the end of the study; it is the values of Y after death that are unobserved, not the survival times. Including death and withdrawal as the reason for their lack of observation is potentially confusing because, unlike the case of withdrawal, values of Y after death are not (according to my definition) missing data unless they are meaningful for analysis. In particular, the outcome in Wu & Carroll’s (1988) application is forced expiratory volume (FEV), which is not meaningful for analysis after death.

Another point of confusion is that the missing values of Y , if considered missing, are not censored in the usual survival analysis sense of having an unknown value beyond a censoring point, but instead are entirely unobserved. The term “informative censoring” is also used in the classical survival analysis setting, where it means something entirely different (see Section 4.4 for details).

Wu & Carroll (1988) did not use the term “informative missingness,” but because the model concerned missing values of Y , it was a small step from their term “informative censoring” to “informative missingness.” This term was indeed used in a similar context by Follman & Wu (1995), Wu & Follman (1998) and Park et al. (2002). My view is that values like blood pressure,

quality of life, or FEV that are censored by death are not examples of informative missingness because they should not be considered missing values.

In Wu & Carroll's (1988) repeated measures model, missingness of Y was assumed to depend on the unobserved values of the random slope and intercept. Because these random effects are not observed, the resulting mechanism is MNAR and hence is nonignorable, according to Rubin's definitions. Informative missingness is often used as a pseudonym for MNAR or nonignorable (Diggle & Kenward 1994) without addressing the parameter distinctness that distinguishes these terms. In the context of missing outcome data, some authors use the term to mean that the distribution of an outcome variable is different for observed and missing cases, perhaps after conditioning on fully observed covariates (e.g., Allen et al. 2003, Higgins et al. 2008). It is unclear how that usage extends to more general patterns of missing data.

In the Statistical Analysis System (2017) manual of data mining, the term "informative missingness" is used to describe a class of models that include indicators for missing values as predictors, a very particular type of MNAR models (Little 2020). This inclusion of missing-data indicators as predictors in imputation is quite common in the machine learning literature.

Given all this semantic confusion, my recommendation is to avoid the terms "informative missingness" and "informative censoring."

4.2. Partial Missing at Random and Ignorability for Parameter Subsets

The definitions of MAR and ignorable missingness apply to the full set of parameters θ in the data model. Write $\theta = (\theta_1, \theta_2)$, where θ_1 and θ_2 are subsets of the components of the model for the data X in Equation 1. Little et al. (2016) define the data as partially MAR for direct likelihood inference about θ_1 , denoted P-MAR(θ_1), if the likelihood (Equation 2) can be factored as

$$L_{\text{full}}(\theta_1, \theta_2, \phi \mid \tilde{x}_{(1)}, \tilde{r}) = L_1(\theta_1 \mid \tilde{x}_{(1)}) \times L_{\text{rest}}(\theta_2, \phi \mid \tilde{x}_{(1)}, \tilde{r}) \text{ for all } \theta_1, \theta_2, \phi, \quad 19.$$

where $L_1(\theta_1 \mid \tilde{x}_{(1)})$ does not involve the model for the missing-data mechanism and $L_{\text{rest}}(\theta_2, \phi \mid \tilde{x}_{(1)}, \tilde{r})$ does not involve the parameters θ_1 . The "partially" in P-MAR(θ_1) refers to the fact that θ_1 can be a subset of θ .

Paralleling Lemma 2, the data are ignorable for direct likelihood inference about θ_1 , denoted IGN(θ_1), if (a) the missing-data mechanism is P-MAR(θ_1) and (b) θ_1 and (θ_2, ϕ) are distinct sets of parameters in the sense defined by Rubin (1976). If the mechanism is P-MAR(θ_1) but θ_1 and (θ_2, ϕ) are not distinct parameters, partial likelihood inference (Cox 1975) based on $L_1(\theta_1 \mid \tilde{x}_{(1)})$ is valid but not fully efficient and might still be adopted to avoid the additional assumptions involved in modeling the missingness mechanism. The partial likelihood $L_1(\theta_1 \mid \tilde{x}_{(1)})$ can also be combined with a prior distribution $\pi_1(\theta_1)$ for θ_1 to obtain a form of pseudo-Bayesian inference, which is not fully Bayes but again avoids the need to model the missingness mechanism. This approach to inference has been proposed and discussed in other contexts (for example, Sinha & Ibrahim 2003, Ventura et al. 2009, Pauli et al. 2011). Little & Zhang (2011) give an application of P-MAR to regression models, selectively discarding data to avoid the need to model a MNAR mechanism.

When $\theta_1 = \theta$, P-MAR and distinctness are weaker than Rubin's (1976) original conditions of MAR and distinctness. The distinctness condition reduces to distinctness between θ and ϕ , as defined in Section 2.2. The P-MAR(θ_1) condition, Equation 19, with $\theta_1 = \theta$ is less restrictive than Rubin's MAR definition, Equation 3, but it does imply the factorization in Equation 4, which is the key condition for validity of inferences about θ based on the ignorable likelihood (Equation 1).

4.3. Missingness as a Coarsening Mechanism

Missing values can be viewed as a form of data coarsening. Heitjan & Rubin (1990), Heitjan (1994), and Jacobsen & Keiding (1995) develop a more general theory for coarsened data that includes heaped, censored, and grouped data as well as missing data.

Denote by $Y = \{y_{ij}\}$ the complete-data matrix in the absence of coarsening, and let $f_Y(y | \theta)$ denote the density of Y under a complete-data model with unknown parameter θ . The observed data, say $y_{ij(1)}$, for each value y_{ij} are that y_{ij} lies in a subset of its sample space Ψ_{ij} , which is determined by a function $y_{ij(1)} = y_{ij(1)}(y_{ij}, c_{ij})$ of y_{ij} and a coarsening variable c_{ij} , subject to the condition that the coarsened subset contains the unobserved true value, that is $y_{ij} \in y_{ij(1)}(y_{ij}, c_{ij})$. For the special case of missing data discussed so far, $C = \{c_{ij}\}$ is simply the matrix of binary missingness indicators, and

$$y_{ij(1)} = \begin{cases} \{y_{ij}\}, & \text{the set consisting of the single true value, if } c_{ij} = 0 \\ \Psi_{ij}, & \text{the sample space of } y_{ij}, \text{ if } c_{ij} = 1 \end{cases}.$$

Uncertainty in the degree of coarsening is modeled by assigning C a probability distribution with conditional density given $Y = y$ equal to $f_{C|Y}(c | y, \phi)$. Write $y = (y_{(0)}, \tilde{y}_{(1)})$ and $c = (c_{(0)}, \tilde{c}_{(1)})$, where $y_{(0)}$ and $\tilde{y}_{(1)}$ are respectively the missing and observed components of Y , and $c_{(0)}$ and $\tilde{c}_{(1)}$ are the missing and observed components of C . The full coarsened-data likelihood is then

$$L_{\text{full}}(\theta, \phi | \tilde{y}_{(1)}, \tilde{c}_{(1)}) = \int \int f_{C|Y}(c_{(0)}, \tilde{c}_{(1)} | y_{(0)}, \tilde{y}_{(1)}, \phi) f_Y(y_{(0)}, \tilde{y}_{(1)} | \theta) dy_{(0)} dc_{(0)},$$

and the likelihood ignoring the coarsening mechanism is

$$L_{\text{ign}}(\theta | \tilde{y}_{(1)}) = \int f_Y(y_{(0)}, \tilde{y}_{(1)} | \theta) dy_{(0)}.$$

The following definitions and lemma generalize the ideas of MAR and ignorable missingness mechanisms to coarsened data:

- Coarsened at random (CAR): The data are CAR at the observed values $y_{(1)} = \tilde{y}_{(1)}$, $c_{(1)} = \tilde{c}_{(1)}$ if

$$f_{C|Y}(c_{(0)}, \tilde{c}_{(1)} | y_{(0)}, \tilde{y}_{(1)}, \phi) = f_{C|Y}(c_{(0)}^*, \tilde{c}_{(1)} | y_{(0)}^*, \tilde{y}_{(1)}, \phi) \text{ for all } c_{(0)}, c_{(0)}^*, y_{(0)}, y_{(0)}^*, \phi.$$

- Ignorable coarsening mechanism: The coarsening mechanism is ignorable if inference for θ based on L_{ign} is equivalent to inference based on the full likelihood L_{full} .

Conditions for ignoring the coarsening mechanism parallel the conditions for ignoring the missingness mechanism. In particular, sufficient conditions for ignoring the coarsening mechanism for likelihood inference at $\tilde{y}_{(1)}$ and $\tilde{c}_{(1)}$ are that (a) the data are CAR, and (b) the parameters θ and ϕ are distinct. Sufficient conditions for ignoring the coarsening mechanism for Bayesian inference are that (a) the data are CAR, and (b) the parameters θ and ϕ have independent prior distributions.

A particular form of coarsening is right-censoring in survival data, and in that setting, “not CAR” is somewhat analogous to the term “informative censoring” in the survival analysis literature, a term that, like “informative missing data,” is not clearly defined. Note that MAR is a special case of CAR, but right-censored data that is CAR is not MAR because missingness depends on the underlying survival time, which is missing for the censored cases.

5. CONCLUSION

With missing data, as a rule there is no such thing as an assumption-free lunch; no analysis of data with missing values, frequentist or model based, is valid for all possible missingness mechanisms, and every methodological article that promises an assumption-free analysis has assumptions lurking somewhere, even if they are not explicitly stated. The only exception to this rule is planned missing data, where a missingness pattern is deliberately created, for example by restricting an expensive measurement to a random subset of units. Thus, the best advice is to not have missing data, or at least to take steps to limit the problem. This is particularly true in areas of statistics where the goal is to limit assumptions, such as randomized clinical trials for assessing medical treatments (see, e.g., NRC 2010).

This review has emphasized the centrality of Rubin's (1976) MAR condition in the analysis of data with missing values, particularly for likelihood-based inference methods. Most existing methods for such data sets assume MAR, and relaxing the MAR assumption trades that assumption for others or explores deviations from MAR by some form of sensitivity analysis. The problem is that we usually cannot tell from the observed data whether or not MAR applies. Rubin himself has argued that with a sufficiently rich set of observed data, MAR is often justified. One empirical example in favor of this view is missing data in the income supplement of the US Current Population Survey, where substantial information is available on income nonrespondents, and a match to an alternative data source suggested that MAR was reasonable. This example is discussed by Little & Rubin (2019, chapter 15). Given this situation, it is important to consider how to make MAR plausible in the design stage of studies by taking steps to limit missing data and collecting covariates that are good predictors of missing values.

For non-likelihood approaches, MAR works in some instances, but stronger conditions such as MCAR or MACAR may be needed, particularly in small samples. My personal view is that Bayesian inference is the best approach to small-sample inference with general patterns of missing data, particularly if MAR can be adequately justified by the availability of good covariates.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I thank the referees for constructive suggestions on an earlier draft.

LITERATURE CITED

- Afifi AA, Elashoff RM. 1966. Missing observations in multivariate statistics 1: review of the literature. *J. Am. Stat. Assoc.* 61:595–604
- Allen AS, Rathouz PJ, Satten GA. 2003. Informative missingness in genetic association studies: case-parent designs. *Am. J. Hum. Genet.* 72:671–80
- Anderson TW. 1957. Maximum likelihood estimates for the multivariate normal distribution when some observations are missing. *J. Am. Stat. Assoc.* 52:200–3
- Barnard J, Frangakis CE, Hill J, Rubin DB. 2002. School choices in NY city: a Bayesian analysis of an imperfect randomized experiment (with discussion). In *Case Studies in Bayesian Statistics*, Vol. 5, ed. R Kass, B Carlin, A Carriquiry, A Gelman, I Verdinelli, M West, pp. 33–97. New York: Springer
- Beesley LJ, Taylor JMG, Little RJ. 2019. Sequential imputation for models with latent variables assuming latent ignorability. *Aust. N. Z. J. Stat.* 61(2):213–33

- Cox DR. 1975. Partial likelihood. *Biometrika* 62(2):269–76
- Diggle P, Kenward MG. 1994. Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc. C* 43:49–73
- De Groot MH, Goel PK. 1980. Estimation of the correlation coefficient from a broken random sample. *Ann. Stat.* 8(2):264–78
- Follman D, Wu M. 1995. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 51(1):151–68
- Harel O, Schafer JL. 2009. Partial and latent ignorability in missing data problems. *Biometrika* 96(1):37–50
- Hartley HO. 1958. Maximum likelihood estimation from incomplete data. *Biometrics* 14:174–94
- Heitjan DF. 1994. Ignorability in general incomplete-data models. *Biometrika* 81(4):701–8
- Heitjan DF. 2004. Estimation with missing data (correspondence). *Biometrics* 50:580
- Heitjan DF, Rubin DB. 1990. Inference from coarse data via multiple imputation with application to age heaping. *J. Am. Stat. Assoc.* 85(410):304–14
- Higgins JPT, White IR, Wood AM. 2008. Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clin. Trials* 5:225–39
- Imbens GW, Rubin DB. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge Univ. Press
- Jacobsen M, Keiding N. 1995. Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Stat.* 23(3):774–86
- Kenward MG, Molenberghs G. 1998. Likelihood based frequentist inference when data are missing at random. *Stat. Sci.* 3(3):236–47
- Lipsitz SR, Ibrahim JG, Zhao LP. 1999. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Am. Stat. Assoc.* 94:1147–60
- Little RJ. 1976. Discussion of “Inference and Missing Data” by D.B. Rubin. *Biometrika* 63:590–92
- Little RJ. 1992. Regression with missing X’s: a review. *J. Am. Stat. Assoc.* 87:1227–37
- Little RJ. 1995. Modeling the drop-out mechanism in longitudinal studies. *J. Am. Stat. Assoc.* 90:1112–21
- Little RJ. 2020. On algorithmic and modeling approaches to imputation in large data sets. *Stat. Sin.* 30:1685–96
- Little RJ, David M. 1988. *Weighting adjustments for non-response in panel surveys*. Tech. Rep., US Dep. Commerce, Bur. Census, Washington, DC
- Little RJ, Rubin DB. 2019. *Statistical Analysis with Missing Data*. New York: Wiley. 3rd ed.
- Little RJ, Rubin DB, Zanganeh SZ. 2016. Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets. *J. Am. Stat. Assoc.* 112:314–20
- Little RJ, Zhang N. 2011. Subsample ignorable likelihood for regression analysis with missing data. *J. R. Stat. Soc. C* 60(4):591–605
- Marini MM, Olsen AR, Rubin DB. 1980. Maximum-likelihood estimation in panel studies with missing data. *Sociol. Methodol.* 11:314–57
- Mealli F, Rubin DB. 2015. Clarifying missing at random and related definitions and implications when coupled with exchangeability. *Biometrika* 102(4):995–1000
- Mealli F, Rubin DB. 2016. Amendments and corrections. *Biometrika* 103(2):491
- NRC (Nat. Res. Counc.). 2010. *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, DC: Nat. Acad. Press
- Park S, Palta M, Shao J, Shen L. 2002. Bias adjustment in analysing longitudinal data with informative missingness. *Stat. Med.* 21:277–91
- Pauli F, Racugno W, Ventura L. 2011. Bayesian composite marginal likelihoods. *Stat. Sin.* 21:149–64
- Pearl J, Glymour M, Jewell NP. 2016. *Causal Inference in Statistics: A Primer*. New York: Wiley
- Peng Y, Little RJ, Raghunathan TE. 2004. An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics* 60:598–607
- Robins JM, Gill RD. 1997. Non-response models for the analysis of non-monotone ignorable missing data. *Stat. Med.* 16:39–56
- Robins JM, Rotnitzky A. 1995. Semiparametric efficiency in multivariate regression models with missing data. *J. Am. Stat. Assoc.* 90:122–29
- Robins JM, Rotnitzky A, Zhao LP. 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* 90:106–21

- Rubin DB. 1976. Inference and missing data. *Biometrika* 63:581–92
- Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6(1):34–58
- Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley
- Scharfstein D, Rotnitzky A, Robins J. 1999. Adjusting for nonignorable dropout using semiparametric models (with discussion). *J. Am. Stat. Assoc.* 94:1096–146
- Seaman S, Galati J, Jackson D, Carlin J. 2013. What is meant by “missing at random?” *Stat. Sci.* 28(2):257–68
- Sinha D, Ibrahim JG. 2003. A Bayesian justification of Cox’s partial likelihood. *Biometrika* 90(3):629–41
- Statistical Analysis System. 2017. *SAS® Visual Data Mining and Machine Learning 8.1 Statistical Procedures*. Cary, NC: SAS Inst. Inc.
- Trawinski IM, Bargmann RW. 1964. Maximum likelihood with incomplete multivariate data. *Ann. Math. Stat.* 35:647–57
- Ventura L, Cabras S, Racugno W. 2009. Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *J. Am. Stat. Assoc.* 104:768–74
- Wu MC, Carroll RJ. 1988. Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* 44:175–88
- Wu MC, Follman DA. 1998. Use of summary measures to adjust for informative missingness in repeated measures data with random effects. *Biometrics* 55(1):75–84



Contents

Modeling Player and Team Performance in Basketball <i>Zachary Turner and Alexander Franks</i>	1
Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What <i>Marc Aerts, Geert Molenberghs, and Olivier Thas</i>	25
Statistical Evaluation of Medical Tests <i>Vanda Inácio, María Xosé Rodríguez-Álvarez, and Pilar Gayoso-Diz</i>	41
Simulation and Analysis Methods for Stochastic Compartmental Epidemic Models <i>Tapiwa Ganyani, Christel Faes, and Niel Hens</i>	69
Missing Data Assumptions <i>Roderick J. Little</i>	89
Consequences of Asking Sensitive Questions in Surveys <i>Ting Yan</i>	109
Synthetic Data <i>Trivellore E. Raghunathan</i>	129
Algorithmic Fairness: Choices, Assumptions, and Definitions <i>Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum</i>	141
Online Learning Algorithms <i>Nicolò Cesa-Bianchi and Francesco Orabona</i>	165
Space-Time Covariance Structures and Models <i>Wanfang Chen, Marc G. Genton, and Ying Sun</i>	191
Extreme Value Analysis for Financial Risk Management <i>Natalia Nolde and Chen Zhou</i>	217
Sparse Structures for Multivariate Extremes <i>Sebastian Engelke and Jevgenijs Ivanovs</i>	241
Compositional Data Analysis <i>Michael Greenacre</i>	271

Distance-Based Statistical Inference <i>Marianti Markatou, Dimitrios Karlis, and Yuxin Ding</i>	301
A Review of Empirical Likelihood <i>Nicole A. Lazar</i>	329
Tensors in Statistics <i>Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu</i>	345
Flexible Models for Complex Data with Applications <i>Christophe Ley, Slađana Babić, and Domien Craens</i>	369
Adaptive Enrichment Designs in Clinical Trials <i>Peter F. Thall</i>	393
Quantile Regression for Survival Data <i>Limin Peng</i>	413
Statistical Applications in Educational Measurement <i>Hua-Hua Chang, Chun Wang, and Susu Zhang</i>	439
Statistical Connectomics <i>Jaewon Chung, Eric Bridgeford, Jesús Arroyo, Benjamin D. Pedigo, Ali Saad-Eldin, Vivek Gopalakrishnan, Liang Xiang, Carey E. Priebe, and Joshua T. Vogelstein</i>	463
Twenty-First-Century Statistical and Computational Challenges in Astrophysics <i>Eric D. Feigelson, Rafael S. de Souza, Emille E.O. Ishida, and Gutti Jogesh Babu</i>	493

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at <http://www.annualreviews.org/errata/statistics>