What is meant by ‘$P(R \mid Y_{obs})$’?

JC Galati

November 28, 2018

Abstract

Missing at Random (MAR) is a central concept in incomplete data methods, and often it is stated as $P(R \mid Y_{obs}, Y_{mis}) = P(R \mid Y_{obs})$. This notation has been used in the literature for more than three decades and has become the de facto standard. In some cases, the notation has been misinterpreted to be a statement about conditional independence. While previous work has sought to clarify the required definitions, a clear explanation of how to interpret the standard notation is lacking. Specifically, it has not been explained clearly why neither $P(R \mid Y_{obs}, Y_{mis})$ nor $P(R \mid Y_{obs})$ are to be interpreted as probability distributions in this statement, and a definition of the function $P(R \mid Y_{obs})$ for non-MAR mechanisms is difficult to locate in the literature. The latter is needed if the standard statement is to provide a useable testable condition. The aim of this paper is to fill these gaps.

Key words and phrases: Missing data, missing at random, ignorability

1 Introduction

In a foundational work on statistical modeling of incomplete data, a joint distribution for the data variables, $Y$, and response indicator variables, $R$, was considered, and conditions under which $R$ need not be modelled were identified [10]. In this work the concept of Missing at Random (MAR) was introduced. A decade later, Little and Rubin stated the condition as follows ([5] p 90; also [6] p 119):

“Observe that if the distribution of the missing-data mechanism does not depend on the missing values $Y_{mis}$, that is, if

$$f(R \mid Y_{obs}, Y_{mis}, \psi) = f(R \mid Y_{obs}, \psi),$$

(5.13)

... Rubin (1976) defines the missing data to be missing at random (MAR) when Eq. (5.13) is satisfied.”

Similar notation was adopted in [12] (p 10): “$P(R \mid Y_{obs}, Y_{mis}, \xi) = P(R \mid Y_{obs}, \xi)$”.

There has been confusion in the literature around how to interpret equation (5.13) correctly, and Mealli and Rubin [7] pointed out that MAR is not a statement about conditional independence. Several factors are likely to have contributed to this confusion. Firstly, two different concepts of ‘missingness’ reside in the standard framework. This has not been identified until recently [4], and if one interprets $Y_{obs}$ to be temporally observed as defined in [4], then the standard notation does state a conditional independence condition. On the other hand, $Y_{obs}$ is intended to be formally observed (see [4]), but neither of the notations ‘$P(R \mid Y_{obs}, Y_{mis})$’ and ‘$P(R \mid Y_{obs})$’ are intended to be interpreted as probability distributions, and this has not been made clear in the literature. Moreover, while the definition of ‘$P(R \mid Y_{obs})$’ is clear for MAR mechanisms, a definition for non-MAR response mechanisms is difficult to locate, and without this definition (5.13) does not give a testable condition.

We explain how $P(R \mid Y_{obs}, Y_{mis})$ and $P(R \mid Y_{obs})$ are to be interpreted in the statement ‘$P(R \mid Y_{obs}, Y_{mis}) = P(R \mid Y_{obs})$’, and we give a simple-to-understand definition
of \( P(R|Y_{obs}) \) for all response mechanisms so that the standard notation can be understood as a testable condition for MAR. We also explain the significance of the function \( P(R|Y_{obs}) \) when the response mechanism is MAR.

## 2 Notation for \((Y, R)\)

Sections 2.1 to 2.5 below summarise notation given in [4] and are included here for completeness.

### 2.1 Random Vectors

Throughout, \( Y \) denotes a random vector modelling the observed and unobserved data comprising all units in the study jointly, and \( R \) denotes a random vector of binary response random variables of the same dimension as \( Y \), where ‘1’ means observed. Joint distributions for the pair of random vectors \((Y, R)\) will be referred to as full distributions.

**Note 2.1.1.** We have no need to treat vectors as denoting column matrices.

**Note 2.1.2.** Typically a data analyst thinks of a given \( y \) as comprising a rectangular matrix with each column pertaining to a specific ‘variable’ (for example, blood pressure) and each row pertaining to a specific unit (for example, an individual in the study). In our notation, the data matrix is shaped so that there is a single row with the data for the various units placed side by side in sets of columns.

### 2.2 Sample Spaces

Let \( R = \{r_1, r_2, \ldots, r_k\} \) be the set of distinct response patterns with \( r_1 = 1 \) denoting the ‘all ones’ vector corresponding to the complete cases. For convenience, let \( r_0 = 0 \) denote the ‘all zeros’ vector corresponding to non-participants, where it may or may not be the case that \( r_j = r_0 \) for some \( j \in \{1, 2, \ldots, k\} \). (We exclude \( j = 0 \) so as to avoid ever having \( P(r_0) = 0 \).) Note that the dot product \( r_j \cdot r_1 \) gives the number of values observed when the \( j^{th} \) response pattern is realised and, in particular, \( r_1 \cdot r_1 \) gives the number of variables in \( R \) (and also in \( Y \)). Let \( \mathcal{Y} = \text{range}(Y) \) be the set of realisable datasets, where a realizable dataset contains complete data including all values that may or may not be observable.

Let \( \Omega = \mathcal{Y} \times R = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_k \) be the full sample space of realisable pairs of datasets and response patterns, where \( \Omega_j = \mathcal{Y} \times \{r_j\} \) for \( r_j \in R \). When the subscript \( j \) of \( r \) is omitted, we denote \( \Omega_j \) by \( \Omega_R \). Let \( \pi_Y \) and \( \pi_R \) denote the projections \((y, r) \mapsto y \) and \((y, r) \mapsto r \), respectively.

Realisations which represent a specific realisable dataset or response pattern only are denoted \( \bar{y} \) and \( \bar{r} \), respectively.

### 2.3 Projections on \( \mathcal{Y} \) and \( \Omega_j \)

For \( j = 1, 2, \ldots, k \), let \( \pi(r_j) : \mathcal{Y} \to \mathcal{Y}^{\pi(r_j)} \) and \( \pi(-r_j) : \mathcal{Y} \to \mathcal{Y}^{\pi(-r_j)} \) denote the projections extracting from each \( y \) vector the vectors of its observed and unobserved values, respectively, according to the response pattern \( r_j \). (In logic, ‘−’ is commonly used for negation.) By convention we set \( \pi(r_0) = \pi(-r_1) = \emptyset \).

To apply these projections correctly over \( \Omega \), we define the following mappings \( o : R \to \{\pi(r) \circ \pi_Y : \Omega_R \to \mathcal{Y}^{\pi(r)}\} \) and \( m : R \to \{\pi(-r) \circ \pi_Y : \Omega_R \to \mathcal{Y}^{\pi(-r)}\} \) and use an abbreviated notation to refer to the images of \((y, r) \in \Omega \) under these mappings:

\[
\begin{align*}
\mathcal{Y}^{o(r)} &:= (y, r)^{o(\pi_Y(y, r))} \\
\mathcal{Y}^{m(r)} &:= (y, r)^{m(\pi_Y(y, r))}.
\end{align*}
\]

(1)

(2)

Additionally, for \( r \in R \) and \( y \in \mathcal{Y} \) set

\[
\mathcal{Y}^{t_o(r)} := \begin{cases} 
\mathcal{Y}^{\pi(r)} & \text{over } \mathcal{Y} \\
(y, r_j)^{o(\pi_Y(y, r))} & \text{over } \Omega
\end{cases}
\]

(3)
and
\[ y_{tm(r)} := \begin{cases} y_{\pi(-r)} & \text{over } Y \\ (y, r_j)^{m(\pi_R(y, r))} & \text{over } \Omega. \end{cases} \tag{4} \]

Note 2.3.1. In the notations in (1)–(4) only the four symbols \( y^{obs(r)}, y^{mi(r)}, y^{to(r)} \) and \( y^{tm(r)} \) are required for working with densities for the distributions for \((Y, R)\) themselves.

Note 2.3.2. The vectors \( y^{obs(r)} \) and \( y^{to(r)} \) have length \( r \cdot r \) while the vectors \( y^{mi(r)} \) and \( y^{tm(r)} \) have length \( r_1 \cdot r_1 - r \cdot r \). Note that these lengths vary from response pattern to response pattern.

2.4 Observable Data Events

Given \((y, r) \in \Omega\), we call
\[ \Omega_{(y, r)} = \{ (y^{obs(r)}, y^{mi(r)}, r) : y_\ast \in Y \} \tag{5} \]
the observed data event corresponding to \((y, r)\). The set \( \Omega_{(y, r)} \) consists of all datasets \( y_\ast \) which have the same observed values as \( y \) (as defined by the response pattern \( r \)). For a fixed \( r \in R \), the events in (5) partition \( \Omega_r \), and over all \( r \) they give a partition of \( \Omega \). These observable data events are the classes of the equivalence relation defined by setting for all \((y_1, r_1), (y_2, r_2) \in \Omega \), 
\((y_1, r_1) \sim_{obs} (y_2, r_2) \) if, and only if, \( r_1 = r_2 \) and \( y_1^{obs(r_1)} = y_2^{obs(r_2)} \).

2.5 Density Functions

We specify full distributions for \((Y, R)\) through density functions \( h : \Omega \rightarrow \mathbb{R} \), with probabilities being determined by integration: 
\[ P(A) = \int_A h \text{ for any } A \subseteq \Omega \text{ for which a probability can be defined (see [1] or [14] for details). Note we suppress the dominating measure in the notation. Two different ways of factorizing } h \text{ are useful:} \]
\[ h(y, r) = f(y) g(r | y) = p(r) p(y | r). \tag{6} \]

The first factorization in (6) is called a selection model factorization of \( h \), and the factor \( g(r | y) \) is called the response mechanism. The second factorization in (6) is called a pattern-mixture factorization, and for each \( r \in R \), we call the conditional density \( p(y | r) \) the pattern-mixture component pertaining to \( r \).

Note 2.6.1. Technically, the symbols \( h, f, g \) and \( p \) denote density functions and 
\( h(y, r), f(y), g(r | y), p(r) \) and \( p(y | r) \) denote real numbers. Because it is common in statistics to use the same symbol to denote different densities, for example a joint density \( f(x_1, x_2) \) and a marginal density \( f(x_1) \), we adopt the usual convention and often refer to density functions by their values.

2.6 Graphical illustration of a full density

Figure 1 provides a pictorial description of a full density and its selection-model and pattern-mixture factorizations. For graphical simplicity the distributions of the \( y \) vectors are depicted as one-dimensional, but in practice these distributions are multi-dimensional.

The marginal probabilities \( p(r_j) \) for the \( k \) response patterns give the marginal distribution for \( R \). These must sum to 1. The marginal density \( f(y) \) is the average of the pattern-mixture densities \( p(y | r_j) \), weighted by their marginal probabilities \( p(r_j) \). The response mechanism evaluated at a fixed \( y \) vector gives the probability distribution for the \( k \) response patterns corresponding to that particular \( y \) vector. These probabilities may vary as \( y \) varies, but for any fixed \( y \) vector they always sum to 1.

The histogram below \( f(y) \) depicts the marginal distribution for \( Y \) that would be observable if the data values could always be observed (that is, if missing data were not possible).
The rectangular regions labelled $\Omega_j$ depict the stratification of $\Omega$ by response pattern. The density below each pattern-mixture component $p(y \mid r_j)$ illustrates the distribution of the $y$ vectors in each stratum. The differing distributional shapes across the pattern-mixture components illustrates the distributional effect of missingness on the complete data (observable and unobservable data values) before any loss of information is incurred due to some data values being unobservable. For example, the shape of the density $p(y \mid r_1)$ depicts the distribution of the complete cases, which in general will differ from the shape of the marginal density $f(y)$. This differing shape explains the potential bias that can result by restricting analyses to complete cases only.

The Full Distribution

\[
p(y \mid r_1) \quad p(y \mid r_2) \quad p(y \mid r_k) \quad f(y)
\]

\[
\Omega_1 \quad \Omega_2 \quad \Omega_k
\]

\[
y^{ob}(r_1) \quad y^{mi}(r_2) \quad \ldots \quad y^{mi}(r_k) \quad y^{ob}(r_2) \quad y^{ob}(r_k)
\]

\[
(y, r_1) \quad (y, r_2) \quad (y, r_k)
\]

\[
g(r_1 \mid y) + g(r_2 \mid y) + \ldots + g(r_k \mid y) = 1
\]

\[
\pi y \quad \uparrow \quad \pi y
\]

\[
\mathbb{R} \quad p(r_1) + p(r_2) + \ldots + c(r_k) = 1
\]

\[
\star r_1 \quad \star r_2 \quad \star r_k
\]

The Marginal Distributions for $R$ and $Y$

Figure 1. Selection-model and pattern-mixture factorisations of a full density.

The vertical bar in each of the sets $\Omega_j$ and in $Y$ represents a single dataset ($y$ vector). In the sets $\Omega_j$, the bar is partitioned into a thick black part representing the values of $y$ that are always observed, and a white unseen part representing the values of $y$ that are never observed (whenever the response pattern $r_j$ is realised). The dotted vertical line in $Y$ represents the fact that $y$ values in the marginal distribution for $Y$ cannot be separated into observed and missing parts, and represent a mixture of observed and missing values averaged over all response patterns. In the figure, and in later notation, we assume that we can re-order the entries of $y$ as is convenient to separate the values into missing and observed parts, and that we know how to reverse this reordering to compare $y$ vectors across the sets $\Omega_j$ and $Y$.

3 Missingness at Random

Missing at Random (MAR) is a property of a response mechanism postulated to hold when the domain of the response mechanism is restricted to a given observable data event, $\Omega(y, r)$. One of these events is illustrated in Figure 2. It consists of all datasets lying between the dotted vertical lines. Note that $y^{ob}(r) = y^{ob}(r)$ is the same for all datasets $y$, in the event, but each $y^{mi}(r)$ vector is different. The definition of MAR was framed by Rubin [10]. Here we state it in a slightly different form.

Definition 3.1. Given $h(y, r) = f(y) g(r \mid y)$ factorised in selection model form together with observed data $\Omega(y, r)$, we say that the response mechanism $g$ is Missing
at Random (MAR) with respect to $\Omega_{(y^*,r)}$ if $g$ is constant on $\Omega_{(y^*,r)}$.

\[ \Omega_r \overset{\leftarrow}{\rightarrow} \Omega_{(y^*,r)} \]

Figure 2. An observable data event in $\Omega_r \subset \Omega$.

**Note 3.1.** MAR as we have defined it is equivalent to the definition framed by Rubin [10], except that we have identified the property as an attribute of the response mechanism of a single full density. Rubin framed the definition for a model of densities $\mathcal{M} = \{ h(\theta, \psi) : (\theta, \psi) \in \Theta \times \Psi \}$, and this can be accommodated simply by requiring that MAR hold with respect to $\Omega_{(y^*,r)}$ for all densities $h(\theta, \psi)$ in $\mathcal{M}$.

Everywhere MAR (in [13]) is accommodated simply by requiring that MAR hold with respect to all observable data events (for all densities in $\mathcal{M}$).

**Note 3.2.** The terminology chosen by Rubin in [10] attributes MAR to the data: “The data ... are missing at random ...”. We have attributed it to the response mechanism because it is possible to have densities $h_1$ and $h_2$ with $h_1$ MAR and $h_2$ not MAR with respect to the same event $\Omega_{(y^*,r)} \subset \Omega$, so the property is not an attribute of the realised data. Nevertheless, Rubin’s terminology makes sense from the following perspective. It can be shown that when MAR holds with respect to $\Omega_{(y^*,r)}$, then $p(y_{mi}^*(r)|y_{ob}(r), r) = f(y_{mi}^*(r)|y_{ob}(r))$ for all $(y_{ob}(r), y_{mi}^*(r), r) \in \Omega_{(y^*,r)}$. This equality says that under MAR, datasets $y_*$ with the same pattern of response $r$ and the same observed values $y_{ob}(r)$ can be drawn ‘independently’ from $\Omega_{(y^*,r)}$.

(see Section 5 of [4], for example).

**Note 3.3.** A pictorial way to interpret the effect of missingness on the distribution of the $y$ values is illustrated in Figure 3.

\[ p(y|r) \propto f(y) g(r|y) \]

Figure 3. Effect of MAR on pattern-mixture component density.

For a fixed response pattern $r$, the response mechanism $g(r|y)$ considered as a function of $y$ alone distorts the shape of the marginal density $f(y)$ to produce the pattern-mixture component $p(y|r)$ (appropriately scaled). The MAR condition requires that when restricted to the observable data event $\Omega_{(y^*,r)}$, the pattern-mixture component and marginal densities have the same shape (but scaled differently).

**Note 3.4.** The main subtlety of the MAR definition surfaces when one attempts to conceive of a response mechanism in which the probability of response pattern $r_*$
given by \( g(r, y) \) varies with the observed values \( y^{\text{obs}}(r) \). Referring back to Figure 1, the equation \( g(r_1 | y) + g(r_2 | y) + \cdots + g(r_k | y) = 1 \) must hold at all times. So \( g(r | y) \) cannot vary in isolation of the probabilities of the other response patterns. There must be one or more other patterns whose probabilities in total vary in the opposite direction to accommodate the change in \( g(r | y) \). If the response mechanism is to be everywhere MAR, then MAR must hold for these offsetting patterns as well, so the probability of the given response pattern cannot depend on any observed variables which are (formally) missing according to these offsetting response patterns. 

4 Why aren’t \( P(R | Y_{\text{obs}}, Y_{\text{mis}}) \) and \( P(R | Y_{\text{obs}}) \) to be treated as probability distributions?

We use an analogy to likelihood theory to explain the reason why \( P(R | Y_{\text{obs}}, Y_{\text{mis}}) \) and \( P(R | Y_{\text{obs}}) \) are not to be treated as probability distributions in the standard notation for MAR.

Consider a model of densities defined on the real numbers \( \mathcal{M} = \{ f_\theta(x) : \theta \in \Theta \} \). There are three different perspectives from which one can view \( \mathcal{M} \). The natural one is to think of \( \Theta \) as the indexing set and each \( f_\theta(x) \) as giving a probability distribution on \( \mathbb{R} \). A second perspective is to consider \( \mathbb{R} \) to be the indexing set and each \( L_\theta(x) = f_\theta(x) \) as giving a likelihood function on \( \Theta \). A third perspective is to consider the entire model as a single function of two variables \( \mathcal{M} : \mathbb{R} \times \Theta \to \mathbb{R} \) sending \((x, \theta)\) to \( f_\theta(x) \).

The same three perspectives apply to the response mechanism \( g(r | y) \). One can consider \( \mathcal{Y} \) to be the indexing set and each \( g(r | y) \) to give a probability distribution for \( R \), one can consider \( \mathcal{R} \) to be the indexing set and each \( g(r | y) \) to give a function on \( \mathcal{Y} \), or one can consider all of the conditional distributions for \( R \) together to give a single function of two vector variables on \( \Omega \). It is the second perspective that is used to interpret the standard notation correctly, not the first. Specifically, the response mechanism is partitioned according to \( \mathcal{R} \) into \( k \) functions:

\[
g_{r_1}(y), g_{r_2}(y), g_{r_3}(y), \ldots, g_{r_k}(y),
\]

where \( g_{r_i}(y) = g(r_i | y) \) for \( i = 1, 2, \ldots, k \). Because the response pattern \( r \) is fixed for each function in (7), for a specific response pattern, \( r \), one can consider the dependence of the corresponding function \( g_r(y) \) separately on \( y^{\text{obs}}(r) \) and \( y^{\text{mis}}(r) \). This is what is intended with the standard notation.

To make the standard notation a testable condition, however, given one of the functions (6) for an arbitrary response mechanism, one must know how to construct a function \( g_r(y^{\text{obs}}(r)) \) for the right hand side of the MAR condition as it is stated in (5.13). It seems difficult to locate such a construction in the literature. We give several different constructions in Section 5 to show how this can be done.

**Note 4.1.** In the same way that a likelihood function does not give a probability distribution, neither should we expect the \( k \) functions in (7) to have meaningful interpretations as probability distributions.

**Note 4.1.** For completeness, we discuss the reasons why \( P(R | Y_{\text{obs}}) \) cannot be interpreted as a conditional probability distribution for \( R \) in any useful way. It is because \( Y_{\text{obs}} \) is a function of \( R \) through the dependence of \( Y = (Y_{\text{obs}}, Y_{\text{mis}}) \) on a specific response pattern \( r \). To interpret \( P(R | Y_{\text{obs}}) \) as a conditional probability distribution for all of \( R \), one must therefore decide on the information about the specific pattern \( r \) that is to be held fixed along with \( Y_{\text{obs}} \) in defining the conditioning event. As the extreme case, one might decide that \( Y_{\text{obs}} \) has meaning only if it is accompanied by the response pattern \( r : (Y_{\text{obs}}, r) \). But this forces \( P(R | Y_{\text{obs}}) = 1 \) if \( R = r \) and \( P(R | Y_{\text{obs}}) = 0 \) otherwise. And attempting to loosen the constraint to allow more response patterns to feature in the distribution with non-zero probabilities similarly does not produce anything useful.


5 Constructions for $P(R \mid Y_{\text{obs}})$

Here we derive two constructions for $P(R \mid Y_{\text{obs}})$. Each suffices to interpret the statement $P(R \mid Y_{\text{obs}}, Y_{\text{mis}}) = P(R \mid Y_{\text{obs}})$ correctly as the MAR condition and works for both realised MAR and everywhere MAR.

Note that for the complete cases, we must have $g_{r_j}(y^{\text{ob}(r_j)}) = g(r_j \mid y)$ because $y^{\text{ob}(r_j)} = y$ by definition. And for MAR response mechanisms and $r_j \neq r_1$, we have no choice but to take $g_{r_j}(y^{\text{ob}(r_j)}) = g(r_j \mid y_*)$ for any $(y_*, r_j) \in \Omega_{(Y^*, r_j)}$ when $g(r_j \mid y)$ is MAR with respect to $\Omega_{(Y^*, r_j)}$. However, we also need a way to define $P(R \mid Y_{\text{obs}})$ for arbitrary $g$ such that the MAR mechanisms can be detected amongst all response mechanisms. In the sections below we give several different constructions.

5.1 A simple construction

Here we give an interpretation for $P(R \mid Y_{\text{obs}})$ that is straightforward to comprehend and easy to communicate. The information that we have available from which to define $P(R \mid Y_{\text{obs}})$ is the set of values $\{ g(r \mid y_*) : (y_*, r) \in \Omega_{(Y^*, r)} \}$. For MAR mechanisms this is a singleton set, and it has at least two values otherwise. An easy way to distinguish between these possibilities is to take

$$g_{r_j}(y^{\text{ob}(r_j)}) = \sup \{ g(r_j \mid y_*) : (y_*, r_j) \in \Omega_{(Y^*, r_j)} \} \quad (8)$$

for $j = 1, 2, \ldots, k$. This is well-defined because $0 \leq g(r \mid y) \leq 1$ for all $(y, r) \in \Omega$. Also, it is clear that with this definition, $g$ is MAR with respect to $\Omega_{(Y^*, r)}$ if, and only if, $g_{r_j}(y_*) = g_{r_j}(y^{\text{ob}(r_j)})$ for all $(y_*, r) \in \Omega_{(Y^*, r)}$.

**Note 5.1.** It makes no difference if we take the infimum instead of the supremum in $(8)$. More generally, we could take various combinations of supremum and infimum in $(8)$ for the different response patterns, so there are many different ways that the notation $P(R \mid Y_{\text{obs}})$ could be defined.

5.2 A construction derived from conditional distributions

An alternative way to define $P(R \mid Y_{\text{obs}})$ which is closer in spirit to the notation can be obtained as follows. For $j = 1, 2, \ldots, k$, consider the $k$ marginal distributions $(Y^{\text{tot}(r_j)}, R)$ of $(Y, R)$ with corresponding densities $h(y^{\text{tot}(r_j)}, r) = \int h(y, r) \, dy^{\text{tot}(r_j)}$. Factorize $h(y^{\text{tot}(r_j)}, r) = f(y^{\text{tot}(r_j)}) \, g(r, y^{\text{tot}(r_j)})$ into marginal and conditional factors and take

$$g_{r_j}^{\text{alt}}(y^{\text{ob}(r_j)}) = \text{the restriction of } \frac{h(y^{\text{tot}(r_j)}, r)}{f(y^{\text{tot}(r_j)})} \text{ to pairs } (y^{\text{ob}(r_j)}, r_j). \quad (9)$$

Note that this is okay because $(y^{\text{ob}(r_j)}, r_j) = (y^{\text{tot}(r_j)}, r)$ when $r = r_j$.

5.3 The significance of $P(R \mid Y_{\text{obs}})$ for MAR response mechanisms

The significance of the function $g_{r}(y^{\text{ob}(y)})$ when $g$ is everywhere MAR is that it is defined on the set $\Omega_{\text{ob}} = \{ y^{\text{ob}(r)}, r : (y, r) \in \Omega \}$ of observable data and it carries all of the information about the dependence of $R$ on $Y$. So, if one had complete information about $f(y)$, then this together with the function $g_{r}(y^{\text{ob}(y)})$ is sufficient to fully reconstruct the entire full distribution for $(Y, R)$. That is, the correct general interpretation of MAR is that no information about the response mechanism is lost through missing data when $g$ is everywhere MAR.

In the literature, one sometimes comes across statements that seem to suggest that distributional information about $f(y)$ is not lost when the response mechanism is everywhere MAR. In general this does not seem to be a valid conclusion. The facts seem to be clear when the only source of missingness is dropout in longitudinal studies. Beyond that however, it is not easy to locate in the literature the correct implications. The most one would seem to be able to concluded in general is that an everywhere MAR response mechanism frees the analyst from having to model $R$ explicitly, and both likelihood theory and multiple imputation rely on the correct...
specification of the full model \( \{ f_\theta(y) : \theta \in \Theta \} \) for the data vector \( Y \) for their validity. That is, these methods work by replacing any information about \( Y \) that is lost through missing data with the assumption that one of the \( f_\theta(y) \) from the model is the correct distribution for \( Y \).

6 Discussion

There has been some confusion in the literature regarding the correct definition(s) of MAR. This is partly because different strength definitions are required for direct likelihood inference compared to frequentist likelihood inference. This was clarified in [13], with the weaker and stronger forms of the definition being called realised and everywhere MAR, respectively (see Section 3 for further details). Additionally, MAR has been misinterpreted in the literature as a form of conditional independence (see [2] page 12, [3] page 381 and [9] page 8, for example). This misinterpretation has been pointed out recently [7], where the authors refer to MAR as being “incorrectly stated or inappropriately redefined in the literature”. This criticism seems unwarranted because the authors provide no explanation as to why this misunderstanding has arisen in the first place.

As discussed in Section 4, to interpret (5.13) as intended by the authors, the reader must be aware that neither \( f(R|Y_{obs}) \) nor \( f(R|Y_{obs},Y_{mis}) \) are being interpreted as probability distributions. In the early textbooks [5, 12] this does not seem to have been pointed out, although something to that effect is stated in [11] (page 50). Additionally, unlike what is the case with the likelihood function, the authors [5, 12] did not change the notation from a conditional distribution to something that would make it clear that \( f(R|Y_{obs},Y_{mis}) \) is not being interpreted as a probability density function. Thirdly, no definition for the right hand side of (5.13) seems to be given in either [5, 6] or [12], and unless one knows how to define this, (5.13) does not give a testable condition for MAR. We suggest that under these circumstances, confusion about the correct interpretation of (5.13) seems quite reasonable.

We have remedied these gaps by explaining carefully why neither \( f(R|Y_{obs},Y_{mis}) \) nor \( f(R|Y_{obs}) \) are to be interpreted as probability distributions in (5.13), and we have shown how the function \( f(R|Y_{obs}) \) can be constructed in an easy-to-understand manner. We hope that filling these gaps will enable readers to approach the literature on statistical methods for incomplete data with a little more confidence and greater understanding.

References

[1] Ash, R.B., and Doléans-Dade, C.A. (2000). Probability & Measure Theory. Academic Press, San Diego, California.
[2] Carpenter, J.R. and M.G Kenward, M.G. (2013). Multiple Imputation and its Application. Wiley, Chichester, West Sussex.
[3] Fitzmaurice, R.M., Laird, N.M. and Were, J.H. (2004). Applied Longitudinal Analysis. Wiley-Interscience, Hobokan, New Jersey.
[4] Galati, J.C. (2018). What does it mean for data to be ‘observed’ or ‘missing’? In preparation.
[5] R.J.A Little and D.B. Rubin, Statistical Analysis with Missing Data. 1st edition, Wiley, New York, 1987.
[6] R.J.A Little and D.B. Rubin, Statistical Analysis with Missing Data. 2nd edition, Wiley-Interscience, Hobokan, New Jersey, 2002.
[7] F. Mealli, D.B. Rubin, Clarifying missing at random and related definitions, and implications when coupled with exchangeability, Biometrika, (2015) 102:4, 995–1000, https://doi.org/10.1093/biomet/asv035; Correction: Biometrika, (2016) 103:2, 491, https://doi.org/10.1093/biomet/asw017.
[8] G. Molenberghs, B. Michials, M.G. Kenward, P.J. Diggle, Monotone missing data and pattern-mixture models, Statistica Neerlandica, (1998) 52, 153–161.
[9] G. Molenberghs, G. Fitzmaurice, M.G. Kenward, A. Tsiatis, G. Verbeke (eds.), Handbook of Missing Data Methodology, Chapman & Hall/CRC, Boca Raton FL, 2015.

[10] D.B. Rubin, *Inference and missing data*, Biometrika, (1976) 63:3, 581–592, https://doi.org/10.1093/biomet/63.3.581.

[11] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[12] J.L. Schafer, *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC, Boca Raton FL, 1997.

[13] S. Seaman, J.C. Galati, D. Jackson, J.B. Carlin, What is Meant by ‘Missing at Random’?, *Statist. Sci.* (2013) 28:2, 257–268, doi:10.1214/13-STS415.

[14] G.R. Shorack, *Probability for Statisticians*, 2nd edition. Springer-Verlag, New York NY, 2000.