Biometrika (2022), **109**, 1, *pp.* 227–241
Printed in Great Britain
doi: 10.1093/biomet/asab002
Advance Access publication 19 May 2021

Missing at random: a stochastic process perspective

BY D. M. FAREWELL, R. M. DANIEL

Division of Population Medicine, School of Medicine, College of Biomedical and Life Sciences, Cardiff University, Cardiff CF14 4YS, U.K.

farewelld@cardiff.ac.uk danielr8@cardiff.ac.uk

AND S. R. SEAMAN

MRC Biostatistics Unit, University of Cambridge, Robinson Way, Cambridge CB2 0SR, U.K.

shaun.seaman@mrc-bsu.cam.ac.uk

SUMMARY

We offer a natural and extensible measure-theoretic treatment of missingness at random. Within the standard missing-data framework, we give a novel characterization of the observed data as a stopping-set sigma algebra. We demonstrate that the usual missingness-at-random conditions are equivalent to requiring particular stochastic processes to be adapted to a set-indexed filtration. These measurability conditions ensure the usual factorization of likelihood ratios. We illustrate how the theory can be extended easily to incorporate explanatory variables, to describe longitudinal data in continuous time, and to admit more general coarsening of observations.

Some key words: Missingness at random; Sigma algebra; Stochastic process.

1. INTRODUCTION

Missing at random (Rubin, 1976) is a central concept in missing-data research. Nevertheless, recent papers (Seaman et al., 2013; Mealli & Rubin, 2015; Doretti et al., 2017) have argued that it remains poorly understood and often inaccurately articulated. The most common formulation (Little & Rubin, 2002, p. 12) is superficially intuitive, but misleading in its detail; accurate formulations exist (Robins & Gill, 1997; Lu & Copas, 2004), but typically hold little heuristic appeal.

For models under which data are missing at random, the likelihood is a product of two terms, a marginal likelihood and a conditional likelihood representing the missingness mechanism. By appealing to the theory of incompletely observed stochastic processes and characterizing these multiplicative components of the likelihood, we give a revealing proof of the likelihood factorization that depends only on simple measurability conditions. These conditions lead to a new, general definition of missingness at random that we find to be both rigorous and intuitive.

We represent data and all conditioning statements in terms of sigma algebras, principally to avoid any confusion over what information is being conditioned upon. As a byproduct, we obtain a single framework for discrete, continuous and more general random variables, whether partially observed in discrete time, in continuous time or under general forms of coarsening.

We tread this path with some trepidation. Rubin described his own initial measure-theoretic treatment of missing data as ‘window dressing’, and then-*Biometrika*-editor David Cox’s advice

was to 'eliminate all that measure theory noise' (Lin et al., 2014). We hope our perspective avoids these pitfalls and exposes a signal that could be overlooked following such noise reduction.

2. Notation

Our starting point is a measurable space $(\Omega, \mathcal{F})$ on which we define probability measures and random variables. We take a pure likelihood perspective (Royall, 1997) and compare just two candidate data-generating models. We aim to assess the evidence, as quantified by their likelihood ratio, in favour of a measure P relative to another measure Q. Since likelihood functions are just pairwise comparisons to an arbitrary reference measure, this simplification is not restrictive.

Following Pollard (2002, pp. 7–11), we adopt de Finetti notation: we allow the set A to also denote the indicator random variable $\mathbb{1}_A$, and reuse the symbol P to mean also its corresponding expectation operator E_P, so that in particular $P(A) = P(\mathbb{1}_A) = E_P(\mathbb{1}_A) = \int \mathbb{1}_A \, \mathrm{d}P$. One way of understanding this broader use of the symbol P is that we are extending its domain from indicator functions to more general random variables. In any event, the main consequence for this paper is that whenever we want to refer to E_P we can simply write P.

The sigma algebra $\mathcal{F}$ is a set of events that represents complete information about the entire stochastic system. In the present context, $\mathcal{F}$ tells us the values of variables that may be observed or missing, and whether they are in fact missing. We avoid the term *complete data*, because its typical usage does not encompass indicators of missingness status, which in our view are certainly data. All available information we simply call *data* and represent by the sigma algebra $\mathcal{D}$. The data sigma algebra $\mathcal{D}$ refers to as-yet-unrealized random variables, and contains all events whose logical status is known once the data are revealed. Until § 3 we remain nebulous about the precise definition of $\mathcal{D}$ but, given our missing-data setting, $\mathcal{D}$ will be a strict subset of $\mathcal{F}$. We write $\sigma(X)$ to denote the sigma algebra generated by a random variable X; X is measurable with respect to the sigma algebra $\mathcal{D}$ if and only if $\sigma(X) \subseteq \mathcal{D}$.

We will assume whenever needed that probability measures are absolutely continuous with respect to one another. Then the customary measure-theoretic definition of the likelihood ratio comparing P and Q is the Radon–Nikodym derivative $\mathrm{d}P/\mathrm{d}Q$ (Andersen et al., 1996, p. 97). Though somewhat formal in appearance, this object is simply a random variable that describes at each point in the sample space Ω the corresponding likelihood ratio of P relative to Q.

In general, the random variable $\mathrm{d}P/\mathrm{d}Q$ will not be $\mathcal{D}$-measurable. This is because P and Q measure the size of each set in $\mathcal{F}$ and, since $\mathcal{F} \supset \mathcal{D}$, the likelihood comparison $\mathrm{d}P/\mathrm{d}Q$ may depend on events whose logical status cannot be determined solely from the data $\mathcal{D}$. In contrast, the likelihood ratio based on the data alone may be represented by its $\mathcal{D}$-measurable restriction

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}} = Q\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \,\middle|\, \mathcal{D}\right). \tag{1}$$

This equality provides some intuition about the meaning of the left-hand side: the conditional expectation, given $\mathcal{D}$ and with respect to Q, of the likelihood ratio $\mathrm{d}P/\mathrm{d}Q$. Loosely, this yields local averages of the random variable $\mathrm{d}P/\mathrm{d}Q$ over each region of the sample space within which the data are constant. As noted by Chang & Pollard (1997, p. 299), this notation is more economical than standard representations such as $\int f \, \mathrm{d}y_{\mathrm{mis}}$ because no y_{mis} need be introduced. Since (1) is data-measurable, it can be used for likelihood comparisons while satisfying standard chain rule relationships between likelihood ratios.

3. Monotone missing data

3.1. *Data*

Throughout this paper, we employ the machinery and methods of stochastic processes. For general missing data, the theory of stochastic processes indexed by sets will be required (Molchanov, 2006); however, we begin with the gentler case of monotone missingness, where it suffices to use standard theory for stochastic processes in discrete time. Unlike some other approaches, the stochastic process perspective permits ideas to be extended from the monotone case to the general setting by making essentially trivial semantic modifications.

Following Rubin (1976), we let $Y = (Y_1, \ldots, Y_n)$ be random variables defined on $(\Omega, \mathcal{F})$, the ranges of which may be any measurable spaces. It is helpful to think of Y as a stochastic process $Y = (Y_m)$ indexed by discrete times $m = 1, \ldots, n$. We observe $Y_1, \ldots, Y_M$, where the integer-valued random variable M is also defined on $(\Omega, \mathcal{F})$ and satisfies $0 \leqslant M \leqslant n$. We do not observe $Y_{M+1}, \ldots, Y_n$: observation of the stochastic process Y ceases at the random time M.

A filtration is a nested family of sigma algebras that captures the idea of information increase over time. When we say that a process (Y_m) is adapted to a filtration $(\mathcal{Y}_m)$, we mean that for each m the random variable Y_m is measurable with respect to the sigma algebra $\mathcal{Y}_m$; adaptedness is just a sequence of measurability conditions. We define $\mathcal{Y}_m = \sigma(Y_l : l \leqslant m)$; the resulting $(\mathcal{Y}_m)$ is called the natural filtration generated by Y, and by construction Y is adapted to $(\mathcal{Y}_m)$.

The random variable M records the time at which observation of Y ceases, and is said to be a stopping time if at each point in time an observer knows whether observation of Y has already ceased. More specifically, M is a stopping time with respect to a filtration $(\mathcal{M}_m)$ if the event $\{M \leqslant m\}$ belongs to $\mathcal{M}_m$ for all m. We can arrange for this to be the case by defining $\mathcal{M}_m = \sigma(\{M \leqslant l\} : l \leqslant m)$. For each m, the sigma algebra $\mathcal{M}_m$ encodes logical information about whether the stopping event has occurred and, if so, at what point it occurred.

Until the stopping time M, we accrue information about both Y and M. We construct a larger filtration $(\mathcal{F}_m)$ by setting $\mathcal{F}_m = \mathcal{Y}_m \vee \mathcal{M}_m$; the notation $\mathcal{Y}_m \vee \mathcal{M}_m$ defines $\mathcal{F}_m$ as the smallest sigma algebra containing both $\mathcal{Y}_m$ and $\mathcal{M}_m$. Informally speaking, $\mathcal{F}_m$ tells us the values of $Y_1, \ldots, Y_m$ and, through knowledge of the indicators $\{M \leqslant 1\}, \ldots, \{M \leqslant m\}$, if and when we have stopped recording measurements before time m. Recall that $\{M \leqslant m\}$ denotes both the subset $\{\omega : M(\omega) \leqslant m\} \subseteq \Omega$ and the indicator random variable $\mathbb{1}_{\{\omega : M(\omega) \leqslant m\}}$.

Information increases until the random time M, at which point no further information is recorded. Since M is also a stopping time with respect to $(\mathcal{F}_m)$, this idea of information increase until a random time may be captured through the elegant definition of the stopping-time sigma algebra $\mathcal{F}_M = \{A \in \mathcal{F} : A \cap \{M \leqslant m\} \in \mathcal{F}_m \text{ for all } m\}$ (Pollard, 2002, pp. 142–43). This $\mathcal{F}_M$ is precisely what we mean by the data sigma algebra $\mathcal{D}$; that is, we define $\mathcal{D} = \mathcal{F}_M$. Despite being decorated with the letter M, we stress that the sigma algebra $\mathcal{F}_M$ is not itself a random object, but consists of that immutable set of events whose logical statuses are always known when the data are revealed, regardless of the particular realized values taken by Y or M.

For simplicity, we shall assume that $\mathcal{F} = \mathcal{F}_n$, so that there are no measurable events beyond those described by Y and M. Similarly, we write $\mathcal{Y} = \mathcal{Y}_n$ or $\mathcal{M} = \mathcal{M}_n$ to describe complete information about Y or M, respectively. As a technical aside, we shall also assume that $(\Omega, \mathcal{F})$ has a product structure that allows measures under which $\mathcal{Y}$ and $\mathcal{M}$ are independent; that is, for all $A \in \mathcal{F}$, there exist $B \in \mathcal{Y}$ and $C \in \mathcal{M}$ such that $A = BC$. In de Finetti notation, the product of events BC connotes $\mathbb{1}_B \mathbb{1}_C = B \cap C$. Recall that two sigma algebras $\mathcal{Y}$ and $\mathcal{M}$ are independent under P if and only if $P(BC) = P(B)P(C)$ for all $B \in \mathcal{Y}$ and $C \in \mathcal{M}$.

It may be valuable at this point to consider the simplest possible example: $n = 1$ with a binary Y_1 that could possibly be missing. In this case Ω is the four-point set $\{00, 01, 10, 11\}$, and $\mathcal{F} = 2^{\Omega}$

is the power set of Ω. For a generic element $bc \in \Omega$, $Y_1(bc) = b$ and $M(bc) = c$. Then $\mathcal{Y}_0$ is the trivial sigma algebra $\sigma(\Omega) = \{\emptyset, \Omega\}$, while $\mathcal{M}_0 = \mathcal{F}_0 = \mathcal{M}_1 = \mathcal{M} = \sigma(\{00, 10\}, \{01, 11\})$; $\mathcal{Y}_1 = \mathcal{Y} = \sigma(\{00, 01\}, \{10, 11\})$, and by assumption $\mathcal{F}_1 = \mathcal{F}$. Rather less obviously, $\mathcal{D} = \mathcal{F}_M = \sigma(\{00, 10\}, \{01\}, \{11\})$. This makes intuitive sense, since when $M = 0$ we cannot hope to distinguish between $Y_1 = 0$ and $Y_1 = 1$. We can check our intuition more formally by verifying that the intersection of each of the atoms $\{00, 10\}$, $\{01\}$ and $\{11\}$ with $\{M \leqslant 0\} = \{00, 10\}$ is in $\mathcal{F}_0$. No finer partition is possible; for example, the singleton set $\{00\}$, when intersected with $\{M \leqslant 0\} = \{00, 10\}$, is not in $\mathcal{F}_0$. Trivially, all intersections of the atoms $\{00, 10\}$, $\{01\}$ and $\{11\}$ with $\{M \leqslant 1\} = \Omega$ belong to $\mathcal{F}_1 = \mathcal{F}$. This simple example is revisited in the Appendix to give concrete illustrations of some of the abstract measure-theoretic quantities deployed later in the paper.

That it might be advantageous to represent incomplete information as a randomly stopped stochastic process was hinted at by Gill et al. (1997), and we too find the simple partition of the sample space provided by $\mathcal{F}_M$ to be both instructive and illuminating as to the nature of the missing-data problem. We stress again that $\mathcal{F}_M$ is not a random object; $\mathcal{D} = \mathcal{F}_M$ is a fixed sub-sigma algebra of $\mathcal{F}$. The sigma algebra $\mathcal{D}$ is strictly smaller than $\mathcal{F}$; that is, there are events in $\mathcal{F}$ that are not elements of $\mathcal{D}$. In particular, $\{00, 01\} \notin \mathcal{D}$ and so $\mathcal{Y} \not\subseteq \mathcal{D}$. Equivalent characterizations of the data sigma algebra are possible, such as $\mathcal{D} = \sigma(M, Y_1, \ldots, Y_M)$, but do not provide such immediate and straightforward conditions for assessing data-measurability as are available once we realize that $\mathcal{D}$ is a stopping-time sigma algebra.

3.2. *Likelihood factorization*

In this section, we describe sufficient conditions for factorization of a likelihood ratio. Gill et al. (1997) and Lu & Copas (2004) investigated when related conditions are also necessary, in the latter case within parametric families of measures. We continue with just the measures P and Q, but, whether comparing many models or just two, it is only when frequentist properties or procedures are of interest that we need concern ourselves with the actual data-generating mechanism. In our pure likelihood context, P and Q may be any pair of measures defined on $(\Omega, \mathcal{F})$, and neither is assumed to be the true data-generating measure (Seaman et al., 2013). This said, the relevance of statistical calculations is clearly enhanced when posited models are plausible reflections of reality, so we examine the evaluation of model assumptions in § 5.

The best available likelihood comparison of P and Q is $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}}$. However, it is customary to suppose that scientific interest in these measures focuses on their behaviour $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}}$ on $\mathcal{Y}$, and that their conditional likelihood ratio given $\mathcal{Y}$, i.e., $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}} = (\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}}/(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}}$ (Hoffman-Jørgensen, 1994, p. 130), is secondary because it describes only the so-called missingness mechanism. Formally, we assume that scientific interest lies in a parameter $\theta : \mathcal{P} \to \Theta$, where $\mathcal{P}$ is a set of probability measures on $(\Omega, \mathcal{F})$, and where $\mathcal{Y}$ is sufficient for θ in the sense that $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}} = 1$ implies $\theta(P) = \theta(Q)$ for any $P, Q \in \mathcal{P}$.

Because scientific interest is restricted to θ and because fully specifying the conditional likelihood ratio $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}}$ can be difficult or inconvenient, we may choose instead to focus evidential comparisons on the marginal likelihood ratio $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}}$. Alas, this marginal likelihood ratio is not data-measurable in general, basically for the same reasons that $\mathrm{d}P/\mathrm{d}Q$ is not data-measurable: because $\mathcal{Y} \not\subseteq \mathcal{D}$. But by analogy with (1), where we express the likelihood ratio $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}}$ in terms of a conditional expectation of $\mathrm{d}P/\mathrm{d}Q$ given the data, we could form a marginal, data-measurable likelihood ratio

$$Q\left(\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{Y}} \,\middle|\, \mathcal{D}\right). \tag{2}$$

A reasonable concern with such a procedure is that the value of (2) may somehow depend on the particular choice of missingness mechanism under P or Q, meaning that (2) was not a marginal likelihood ratio for θ after all. We show in Lemma 2 that within certain equivalence classes no such dependence exists, and hence within these equivalence classes (2) may justifiably be called a marginal likelihood ratio. This result is based on a foundational lemma strongly related to the exchange of *seeing* and *doing* in causal inference (Pearl, 2009), which we state first.

LEMMA 1. *The likelihood ratio* $\mathrm{d}P/\mathrm{d}Q$ *is* $\mathcal{D}$*-measurable if and only if* $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{D}} = 1$.

The proof is direct: $\mathrm{d}P/\mathrm{d}Q$ is $\mathcal{D}$-measurable if and only if $\mathrm{d}P/\mathrm{d}Q = (\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}}$, and hence if and only if $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{D}} = 1$. The implication is that when $\mathrm{d}P/\mathrm{d}Q$ is $\mathcal{D}$-measurable, expectations conditional on $\mathcal{D}$ may be taken interchangeably with respect to P or Q; that is, $P(A \mid \mathcal{D}) = Q(A \mid \mathcal{D})$ for all $A \in \mathcal{F}$.

LEMMA 2. *Write* $P \sim Q \pmod{\mathcal{Y}}$ *if and only if* $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}}$ *is* $\mathcal{D}$*-measurable. Let* P, P', Q *and* Q' *be measures all belonging to the same* $\sim$*-equivalence class, and suppose that* $(\mathrm{d}P'/\mathrm{d}P)|_{\mathcal{Y}} = (\mathrm{d}Q'/\mathrm{d}Q)|_{\mathcal{Y}} = 1$. *Then*

$$Q\left(\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{Y}} \,\middle|\, \mathcal{D}\right) = Q'\left(\left.\frac{\mathrm{d}P'}{\mathrm{d}Q'}\right|_{\mathcal{Y}} \,\middle|\, \mathcal{D}\right) = 1/P\left(\left.\frac{\mathrm{d}Q}{\mathrm{d}P}\right|_{\mathcal{Y}} \,\middle|\, \mathcal{D}\right).$$

To prove the first equality, we have only to exchange $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}}$ for $(\mathrm{d}P'/\mathrm{d}Q')|_{\mathcal{Y}}$; then, because $(\mathrm{d}Q'/\mathrm{d}Q)|_{\mathcal{Y}} = 1$, which is clearly $\mathcal{D}$-measurable, and because $(\mathrm{d}Q'/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}}$ is $\mathcal{D}$-measurable since $Q \sim Q' \pmod{\mathcal{Y}}$, it follows that $\mathrm{d}Q'/\mathrm{d}Q$ is itself $\mathcal{D}$-measurable and that, by Lemma 1, expectations conditional on $\mathcal{D}$ can be taken interchangeably with respect to Q or Q'. The proof of the second equality depends critically on $P \sim Q \pmod{\mathcal{Y}}$, but is otherwise unilluminating, so we omit it. Nevertheless, this second equality describes an important symmetry property exhibited by likelihood ratios, but not necessarily by modifications thereof like (2).

Within an equivalence class, Lemma 2 says that for any P and P' that agree on $\mathcal{Y}$, it does not matter whether we use P or P' in the numerator of the marginal likelihood ratio (2) and, so long as Q and Q' agree on $\mathcal{Y}$, we may use either Q or Q' in the denominator; the value of (2) is unchanged. Consequently, within an equivalence class, calling (2) a marginal likelihood ratio appears justifiable. Conversely, if $P \not\sim Q \pmod{\mathcal{Y}}$, then (2) is not a marginal likelihood ratio: even the basic symmetry property of Lemma 2 fails to hold. So the restriction of likelihood comparisons to measures in the same equivalence class is important. Anticipating slightly, one equivalence class will be the set of measures under which data are missing at random.

Another reason that (2) may justifiably be called a marginal likelihood ratio is that within an equivalence class the usual likelihood ratio $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}}$ includes (2) as a multiplicative factor. This fact is so important that we state it as a lemma.

LEMMA 3. *If* $P \sim Q \pmod{\mathcal{Y}}$, *then*

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}} = Q\left(\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{Y}} \,\middle|\, \mathcal{D}\right) \times \left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{F}|\mathcal{Y}}.$$

The result follows directly from the fact that $P \sim Q \pmod{\mathcal{Y}}$ if and only if $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}}$ is $\mathcal{D}$-measurable, and the latter can therefore be brought outside the conditional expectation.

Lemmas 2 and 3 show that, given a pair of measures defined only on the restricted space $(\Omega, \mathcal{Y})$, the marginal likelihood ratio (2) is a multiplicative factor in the likelihood ratio formed

by extending these measures to the whole of $(\Omega, \mathcal{F})$ and restricting to $\mathcal{D}$. Moreover, the marginal likelihood ratio (2) is invariant with respect to the particular way these extensions are constructed, always provided that such extended measures belong to the same $\sim$-equivalence class. The downside to this attractive invariance is that we must choose an equivalence class within which to work. Moreover, because the equivalence classes are defined by a data-measurability condition, there can be no evidence in the data to support one choice over another (see Molenberghs et al., 2008). This choice must be made based on convenience, on meta-data considerations or, most usually, on a combination of both. We discuss convenience first, before moving on to meta-data considerations later in the paper.

It is fairly clear that computing the marginal likelihood ratio (2) would be reasonably straightforward if in fact $\mathcal{Y}$ and $\mathcal{M}$ were independent under P and Q, and this may be one reason to prefer equivalence classes in which such measures appear. In fact, the next lemma shows that all independence measures belong to the same equivalence class, within which the marginal likelihood ratio for θ has a very simple form.

LEMMA 4. *Suppose that $\mathcal{Y}$ and $\mathcal{M}$ are independent under P and Q. Then $P \sim Q \pmod{\mathcal{Y}}$ and*

$$Q\left(\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{Y}} \;\middle|\; \mathcal{D}\right) = \left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}|\mathcal{M}}.$$

To prove this, we use the fact that independence of $\mathcal{Y}$ and $\mathcal{M}$ under P and Q means that $\mathrm{d}P/\mathrm{d}Q = (\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}} \times (\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{M}}$, and hence that $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}} = (\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{M}}$, which is certainly $\mathcal{D}$-measurable. Conversely, $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}} = (\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{M}}$ under independence, so $Q\{(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}} \mid \mathcal{D}\} = Q\{(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}} \mid \mathcal{D}\}/(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{M}}$, as required.

The equivalence class in which the independence measures lie is precisely the class of measures satisfying the condition Rubin (1976) calls missingness at random. Lemma 2 tells us that within this class, the marginal likelihood ratio depends only on the behaviour of P and Q on $\mathcal{Y}$, so using measures under which $\mathcal{Y}$ and $\mathcal{M}$ are independent may be a convenient way to compute it. Lemma 3 shows that the marginal likelihood ratio for θ is a factor in the full likelihood, and Lemma 4 gives us the form of the marginal likelihood ratio directly. We put all these pieces together in the following central result.

THEOREM 1. *Let P, P', Q and Q' be measures such that $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{Y}} = (\mathrm{d}Q/\mathrm{d}Q')|_{\mathcal{Y}} = 1$. Suppose that $\mathcal{Y}$ and $\mathcal{M}$ are independent under P' and Q', and further assume that both $P \sim P' \pmod{\mathcal{Y}}$ and $Q \sim Q' \pmod{\mathcal{Y}}$. Then*

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}} = \left.\frac{\mathrm{d}P'}{\mathrm{d}Q'}\right|_{\mathcal{D}|\mathcal{M}} \times \left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{F}|\mathcal{Y}}.$$

Independence measures P' and Q' exist because of the product structure on $(\Omega, \mathcal{F})$. The proof depends on Lemma 4, from which we deduce that $P' \sim Q' \pmod{\mathcal{Y}}$, and therefore also $P \sim Q \pmod{\mathcal{Y}}$ by transitivity. Since $P \sim Q \pmod{\mathcal{Y}}$, Lemma 3 provides a factorization of the full likelihood ratio, and Lemma 4 gives us the simpler form of the first term, as required.

Reducing likelihood factorizations to questions of data-measurability is a very general idea, and in particular none of the current section depends in any way on monotonicity of missing data. Indeed, our general approach applies equally well in any setting where the structure of the observed data is not fixed in advance. Unobserved quantities need not be thought of as data, but could simply be latent variables, for example random effects. This is the perspective taken by

Farewell et al. (2017). An appealing advantage of the theory developed in the present paper is that it can more easily be extended to cases where the range of M is uncountable (Commenges & Gegout-Petit, 2015, p. 14), and we explore some of these possibilities in § 5. Gill et al. (1997) provided an exactly analogous likelihood decomposition in the setting of coarsened data, and they too highlighted the simplification arising from the working assumption of independence, which Jacobsen & Keiding (1995) call a reference model.

So far we have shown that a working independence conditional likelihood ratio is the unique marginal likelihood ratio for the parameter of interest θ within a particular equivalence class, and that this in turn is a multiplicative factor in the full likelihood ratio. It remains to demonstrate that missingness at random fully characterizes this equivalence class, and to explore the suitability of this class for use in substantive problems of statistical inference.

3.3. *Measurability*

The factorization of Theorem 1 holds whenever the conditional likelihood ratios $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{F}|\mathcal{Y}}$ and $(\mathrm{d}Q/\mathrm{d}Q')|_{\mathcal{F}|\mathcal{Y}}$ are $\mathcal{D}$-measurable; we now discuss how to determine whether they are. We assume throughout this section that P, P', Q and Q' satisfy the conditions of Theorem 1; that is, P and P' agree on $\mathcal{Y}$, Q and Q' agree on $\mathcal{Y}$, and $\mathcal{Y}$ and $\mathcal{M}$ are independent under P' and Q'. We focus on the relationship between P and P', with analogous considerations needed for Q and Q'. It seems likely that the conditional likelihood ratio $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{F}|\mathcal{Y}}$ should be expressible in terms of conditional densities of M given $\mathcal{Y}$ under P and P', and this we now assert to be the case. The proof of Lemma 5 is not especially illuminating, so we defer it to the Appendix.

LEMMA 5. *Let P and P' be measures on $(\Omega, \mathcal{F})$. Define the stochastic processes (p_m) and (p'_m) by $p_m = P(M = m \mid \mathcal{Y})$ and $p'_m = P'(M = m \mid \mathcal{Y})$. Then $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{F}|\mathcal{Y}} = p_M/p'_M$, the ratio of these stochastic processes evaluated at the random stopping time M.*

We deduce that $P \sim P' \pmod{\mathcal{Y}}$ if both p_M and p'_M are $\mathcal{D}$-measurable. We see immediately that the latter is always $\mathcal{D}$-measurable: because $\mathcal{Y}$ and $\mathcal{M}$ are independent under P', $p'_m = P'(M = m \mid \mathcal{Y}) = P'(M = m)$, and hence the process (p'_m) is in fact a deterministic sequence. The value of p'_M then depends only on M, which is certainly $\mathcal{D}$-measurable.

Data-measurability of p_M is more subtle and not guaranteed. Our next result introduces our definition of missingness at random, and proves that this condition on a measure P establishes that the corresponding p_M is data-measurable.

LEMMA 6. *Suppose that $P(M = m \mid \mathcal{Y}) = P(M = m \mid \mathcal{Y}_m)$ for every m. Then (p_m) is adapted to $(\mathcal{F}_m)$, and p_M is $\mathcal{F}_M$-measurable.*

The proof is not difficult. For all m, the condition $P(M = m \mid \mathcal{Y}) = P(M = m \mid \mathcal{Y}_m)$ ensures that $P(M = m \mid \mathcal{Y})$ is $\mathcal{Y}_m$-measurable, i.e., (p_m) is adapted to $(\mathcal{Y}_m)$. Since $\mathcal{Y}_m \subseteq \mathcal{F}_m$, the stochastic process (p_m) must also be adapted to $(\mathcal{F}_m)$. Because M is an $(\mathcal{F}_m)$-stopping time, standard stochastic process results allow us to conclude that p_M is $\mathcal{F}_M$-measurable, as required.

We make several comments. First, the condition of Lemma 6 is equivalent to the rigorous, everywhere version of missingness at random (Seaman et al., 2013), but is simpler to state: by conditioning on sigma algebras $\mathcal{Y}_m$, we equate random variables, not a large set of conditional probabilities. Second, it applies equally well to a random variable Y taking values in uncountable spaces; neither discretization (Seaman et al., 2013) nor conditioning on a set of measure zero is required. Third, there is a striking visual similarity of the condition $P(M = m \mid \mathcal{Y}) = P(M = m \mid \mathcal{Y}_m)$ for all m to the ubiquitous, informal missing-at-random definition $P(M \mid Y) = P(M \mid Y_{\mathrm{obs}})$ (Little & Rubin, 2002, p. 12). Interpreted literally, though, the two definitions have

rather different meanings: our definition of missing at random refers to a sequence of fixed values m, and not to the random variables M and Y_{obs}. The measure-theoretic perspective encourages us to distinguish sharply between the random variable M and a generic realized value m, while the Y_{obs} notation blurs this distinction. In our view, this blurring, and the apparent collapse of multiple conditions into one that results from it, leads to much of the confusion surrounding the Y_{obs} and Y_{mis} notation (Seaman et al., 2013). This is another reason we think it is helpful to instead understand missingness at random as an adaptedness condition.

We have shown that if $P(M = m \mid \mathcal{Y}) = P(M = m \mid \mathcal{Y}_m)$ for every m, then p_M and $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{F}|\mathcal{Y}}$ are $\mathcal{D}$-measurable, and hence P belongs to the same $\sim$-equivalence class as the independence measures. If the same is true for the measure Q, then the conditions of Theorem 1 are met, and the likelihood factorization follows. We now proceed to extend this result to nonmonotone missing data and, ultimately, to more general forms of missingness.

4. NONMONOTONE MISSING DATA

We turn now to the general case, where there need be no natural ordering of the components of Y; the observations may be obtained simultaneously or in an arbitrary order, and any possible subset of the variables may be observed. Despite this generality, remarkably few notational changes are needed from the ordered, monotone case; we simply reinterpret what we have written to this point in terms of stochastic processes indexed by sets (Molchanov, 2006, p. 334). Our subscript m becomes a set, so that if $m = \{1, 3, 4\}$ then $Y_m = (Y_1, Y_3, Y_4)$. The most important change from the monotone case is that we now understand M as a random subset of $\{1, \ldots, n\}$, representing the subset of variables that are observed. There is no total ordering of the subsets of $\{1, \ldots, n\}$, but we exploit the partial ordering given by set inclusion and interpret $l \leqslant m$ as $l \subseteq m$, which describes a lattice of subsets on which stochastic processes may be defined. Once again, we stop observing Y at the random set M on this lattice, but now there are potentially multiple routes by which we may arrive at a given point. Just as before, however, we observe the values of all random variables Y_m for which $m \leqslant M$, i.e., for which $m \subseteq M$.

We define $\mathcal{Y}_m = \sigma(Y_l : l \leqslant m)$, $\mathcal{M}_m = \sigma(\{M \leqslant l\} : l \leqslant m)$ and $\mathcal{F}_m = \mathcal{Y}_m \vee \mathcal{M}_m$ just as in the monotone case, where now $(\mathcal{F}_m)$ is a set-indexed filtration. As before, $\mathcal{D} = \mathcal{F}_M$, now a stopping-set sigma algebra. The definitions of the probability measures P, P', Q and Q' are unaltered, and likelihood factorization again boils down to $\mathcal{F}_M$-measurability of p_M, where $p_m = P(M = m \mid \mathcal{Y})$, with similar considerations needed for Q. The missing-at-random condition is unaltered and forms the premise of our central theorem, which we now state formally.

THEOREM 2. *Let P, P', Q and Q' be measures such that $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{Y}} = (\mathrm{d}Q/\mathrm{d}Q')|_{\mathcal{Y}} = 1$. Suppose that $\mathcal{Y}$ and $\mathcal{M}$ are independent under P' and Q'. Further, assume that $P(M = m \mid \mathcal{Y}) = P(M = m \mid \mathcal{Y}_m)$ and $Q(M = m \mid \mathcal{Y}) = Q(M = m \mid \mathcal{Y}_m)$ for all m. Then $P \sim P' \sim Q' \sim Q \pmod{\mathcal{Y}}$, and*

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}} = \left.\frac{\mathrm{d}P'}{\mathrm{d}Q'}\right|_{\mathcal{D}|\mathcal{M}} \times \left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{F}|\mathcal{Y}}.$$

The proof is identical to that in the monotone case: adaptedness of the processes $P(M = m \mid \mathcal{Y})$ and $Q(M = m \mid \mathcal{Y})$ to the set-indexed filtration $(\mathcal{F}_m)$, and the fact that M is an $(\mathcal{F}_m)$-stopping set ensures that $P \sim P' \pmod{\mathcal{Y}}$ and $Q \sim Q' \pmod{\mathcal{Y}}$. Even in this unordered setting, the stochastic process techniques provide us with a direct proof of data-measurability of the likelihood ratios

$(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{F}|\mathcal{Y}}$ and $(\mathrm{d}Q/\mathrm{d}Q')|_{\mathcal{F}|\mathcal{Y}}$. The final likelihood factorization follows from Theorem 1, which was already of sufficient generality to accommodate the set-indexed case.

The likelihood factorization depends on the adaptedness of stochastic processes, so it is worth considering our ability to assess this collection of conditions. The lattice structure implicit in this formulation is reminiscent of the randomized monotone missingness mechanisms of Robins & Gill (1997), wherein future observation can depend on previous measurements within the history of a particular branch. But, as noted by Robins & Gill (1997), more complicated dependence structures are also possible. It may sometimes be appropriate to adopt the missingness-at-random assumption openly, but uncritically, and to use its simple $\sim$-equivalence class to compare likelihoods or conduct inference under this working assumption. Alternatively, perhaps with improved intuition about what it means for a process to be adapted to a set-indexed filtration, the plausibility of the missingness-at-random assumption may be directly and critically assessed even in cases of nonmonotone missing data. Failing this, a fastidious analyst must abandon the appealing generality of missing at random, and assess instead conditions that are stronger and more easily assessed, such as randomized monotone missingness (Robins & Gill, 1997) or stability (Farewell et al., 2017). In § 5 we give specific examples of such stronger conditions.

5. Extensions

5.1. *Modified notation*

We now offer three extensions to the classical setting of Rubin (1976), showing how stochastic process theory adapts naturally to situations where traditional approaches can be cumbersome.

In this more general setting, we align our notation with standard choices made in the study of stochastic processes. Let $Y = (Y_t : t \in T)$ be a stochastic process indexed by a set T, where the latter may be uncountable, but is equipped with a partial or total ordering. We observe Y_t for all $t \leqslant \tau$, where the observed random variable τ takes values in T; we do not observe any Y_t for $t \not\leqslant \tau$. The process Y is adapted to its natural filtration $(\mathcal{Y}_t)$, and again τ is a stopping time or more general stopping object with respect to the filtration $(\mathcal{T}_t)$ generated by the process $(\{\tau \leqslant t\})$. We assume that $\mathcal{F} = \mathcal{Y} \vee \mathcal{T}$, where $\mathcal{Y} = \bigvee_t \mathcal{Y}_t$ and similarly $\mathcal{T} = \bigvee_t \mathcal{T}_t$. We define another filtration $(\mathcal{F}_t)$ by $\mathcal{F}_t = \mathcal{Y}_t \vee \mathcal{T}_t$. The observed data are then given by the generalized stopping-time sigma algebra $\mathcal{D} = \mathcal{F}_\tau = \{A \in \mathcal{F} : A \cap \{\tau \leqslant t\} \in \mathcal{F}_t \text{ for all } t\}$.

In this more general setting, it may well be the case that $P(\tau = t \mid \mathcal{Y}) = 0$ for all possible values of t, for example because τ might be a continuous random variable. Consequently, a slightly modified version of the missing-at-random condition will be required. As in the proof of Lemma 5, we could still define the conditional law μ of τ satisfying $\mu(D, \cdot) = P\{\tau^{-1}(D) \mid \mathcal{Y}\}$ and conditional densities like $p_t = (\mathrm{d}\mu/\mathrm{d}\nu)(t, \cdot)$ with respect to some dominating measure ν, but to avoid these abstractions we work instead with more familiar objects $P_t = P(\tau \leqslant t \mid \mathcal{Y})$ that are analogous to conditional cumulative distribution functions. Like $(p_t : t \in T)$, the collection $(P_t : t \in T)$ is a stochastic process on the partially ordered set T. Since (p_t) is adapted if and only if (P_t) is adapted, it then suffices to define missingness at random in general as follows.

Lemma 7. *If $P(\tau \leqslant t \mid \mathcal{Y}) = P(\tau \leqslant t \mid \mathcal{Y}_t)$ for every t, then (P_t) is adapted to $(\mathcal{F}_t)$, and $P \sim P' \pmod{\mathcal{Y}}$ for any measure P' under which $\mathcal{Y}$ and $\mathcal{M}$ are independent.*

5.2. *Explanatory variables*

We suppose there is now some fully observed covariate information $\mathcal{X} \subseteq \mathcal{D}$ available and that our interest is in the conditional likelihood ratio $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}|\mathcal{X}}$, as might be the case in a

regression modelling context. An immediate advantage of the use of sigma algebras is that we can simply absorb this covariate information into the two existing filtrations $(\mathcal{Y}_t)$ and $(\mathcal{T}_t)$, and assume that for all t we have $\mathcal{X} \subseteq \mathcal{Y}_t$ and $\mathcal{X} \subseteq \mathcal{T}_t$. If $\mathcal{P} = \{P_{\alpha\beta\gamma}\}$ is a family of models wherein α, β and γ respectively characterize behaviour on $\mathcal{X}$, regression coefficients and distributions of residuals, with $\theta(P_{\alpha\beta\gamma}) = \beta$, then $(\mathcal{Y} \mid \mathcal{X})$ is sufficient for θ in the sense that for any $P, Q \in \mathcal{P}$, $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}|\mathcal{X}} = 1$ implies $\theta(P) = \theta(Q)$.

Lemma 4 will now hold under the weaker condition that $\mathcal{Y}$ and $\mathcal{T}$ are conditionally independent given $\mathcal{X}$. By incorporating $\mathcal{X}$ into the two existing filtrations, we arrange for the missing-at-random condition to remain notationally unchanged: $P(\tau \leqslant t \mid \mathcal{Y}) = P(\tau \leqslant t \mid \mathcal{Y}_t)$ for all t, recalling that now $\mathcal{X} \subseteq \mathcal{Y}_t$ for all t. Like the condition employed by Sweeting et al. (2010), this allows dependence of the missingness mechanism on covariates and, to the extent permitted by data-measurabililty, on Y. This could be called covariate-dependent missingness at random, but, as noted by Hedeker & Gibbons (1997), versions such as this should be carefully distinguished from similarly named alternatives, where missingness depends only on covariates and not on observed responses (Little, 1995). The relevant likelihood factorization is

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}|\mathcal{X}} = \left.\frac{\mathrm{d}P'}{\mathrm{d}Q'}\right|_{\mathcal{D}|\mathcal{T}} \times \left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{F}|\mathcal{Y}},$$

where $Q\{(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}|\mathcal{X}} \mid \mathcal{D}\} = (\mathrm{d}P'/\mathrm{d}Q')|_{\mathcal{D}|\mathcal{T}}$ is the marginal likelihood particular to this equivalence class, integrating over the missing data. Since $\mathcal{X} \subseteq \mathcal{T}$, this likelihood is also conditional on $\mathcal{X}$. Our insistence that $\mathcal{X} \subseteq \mathcal{Y}$ pays dividends not only in notational brevity, but also in ensuring that intuitively important terms such as $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{Y}|\mathcal{X}}$ remain well-defined.

The ability to include such covariate information $\mathcal{X}$ is hugely important in missing-data considerations and, more broadly, in matters of causal inference. While hardly ever a trivial exercise, building a set of always-available data $\mathcal{X}$ such that the required adaptedness condition may plausibly be assumed to hold is sometimes more straightforward than exerting external control of the processes that lead to aspects of Y going unobserved.

5.3. *Longitudinal data*

Consider the case where $Y = (Y_u)_{u \geqslant 0}$ is a continuous-time stochastic process and $\tau \subseteq [0, \infty)$ describes the finite set on which Y is observed. Such a construction describes unbalanced longitudinal data (Diggle et al., 2002, p. 282), where each subject gives rise to a random number of observations and these may be recorded at arbitrary points in time. For example, an individual observed at $u = 0.2$, $u = 0.5$ and $u = 0.8$ would have $\tau = \{0.2, 0.5, 0.8\}$. The corresponding filtration $(\mathcal{Y}_t : t \in T)$ is indexed by the power set $T = 2^{[0,\infty)}$ and therefore not totally ordered; instead, we again have a partial ordering $s \leqslant t$ of the sets s and t if and only if $s \subseteq t$, and thus $\mathcal{Y}_s \subseteq \mathcal{Y}_t$. The set T of all possible subsets of $[0, \infty)$ is uncountable, so our modified version of the missing-at-random condition will be required: $P(\tau \leqslant t \mid \mathcal{Y}) = P(\tau \leqslant t \mid \mathcal{Y}_t)$ for all $t \in T$. An example set $t \in T$ for which such probabilities might be nonzero is $t = [0, 0.3] \cup [0.4, 0.9]$, in which case $\tau = \{0.2, 0.5, 0.8\} \leqslant t$.

To our knowledge, characterizations of missingness at random for general longitudinal data are rare, and it is worth considering again our ability to assess this condition. Especially in longitudinal settings, the causal processes that lead to τ taking on any particular value rarely operate by first obtaining all possible information $\mathcal{Y}$ about Y and then deciding what subset of this information to reveal. Nevertheless, our formulation of missingness at random implicitly invites us to evaluate directly whether each $P(\tau \leqslant t \mid \mathcal{Y})$ in fact depends only on $\mathcal{Y}_t$.

We suggest taking a dynamic approach, in which the stopping set τ is itself thought of as arising from a set-valued stochastic process $(\tau_u)_{u\geqslant 0}$, with the set $\tau_u = \tau \cap [0, u]$ defined to be the observed subset of τ up to and including time u. At each time u we observe a set-valued increment $\mathrm{d}\tau_u$ in this process, where $\mathrm{d}\tau_u = \tau_u \setminus \tau_{u-}$ and $\tau_{u-} = \tau \cap [0, u)$. When no observation is made at time u, the increment $\mathrm{d}\tau_u$ is the empty set; when we make an observation at time u, the increment is the singleton set $\{u\}$.

With this more dynamic perspective, we are in a better position to assess the plausibility of missingness at random. Consider, for example, a patient under a so-called doctor's care regime, which Grüger et al. (1991) define to mean that future examination times are determined entirely on the basis of earlier observations. Under such a regime, for an arbitrary set $t \in T$ we can decompose $P(\tau \leqslant t \mid \mathcal{Y})$ using the product integral

$$\prod_{u\in[0,\infty)} P(\mathrm{d}\tau_u \leqslant t \mid \tau_{u-} \leqslant t, \mathcal{Y}),$$

wherein $\mathrm{d}\tau_u, t$ and τ_{u-} are all sets. Whenever $u \in t$, the integrand is unity; elsewhere it specifies the instantaneous probability that no observation is made at time u, given $\mathcal{Y}$, and given the fact that all observation times to date lie in the set t. But we have asserted that under the doctor's care scenario, future observation times are determined only with reference to past observations $\mathcal{Y}_{\tau_{u-}}$, which must be a subset of $\mathcal{Y}_t$ since we are conditioning on the event $\tau_{u-} \leqslant t$. Hence $P(\mathrm{d}\tau_u \leqslant t \mid \tau_{u-} \leqslant t, \mathcal{Y}) = P(\mathrm{d}\tau_u \leqslant t \mid \tau_{u-} \leqslant t, \mathcal{Y}_t)$, and our decomposition of $P(\tau \leqslant t \mid \mathcal{Y})$ multiplies back up to give $P(\tau \leqslant t \mid \mathcal{Y}_t)$, as required to establish that missingness at random holds. Here $P(\mathrm{d}\tau_u \leqslant t \mid \tau_{u-} \leqslant t, \mathcal{Y})$ is a kind of intensity process, and our doctor's care assumption is very like independent censoring (Andersen et al., 1996, p. 139).

In situations where factors outside the doctor's control may influence the number and timing of observations, such considerations become more delicate. For instance, if imperfectly measured subject-specific quantities such as a participant's overall health may influence both Y and τ, the missing-at-random condition will not in general be satisfied. Farewell et al. (2017) make use of causal directed acyclic graphs to help determine if the likelihood contribution of the random observation times τ may safely be ignored.

5.4. *Coarsened observations*

For general coarsening of observations (Heitjan & Rubin, 1991), we shall assume as usual that $\mathcal{Y}$ represents complete information about some random variable Y. As in the longitudinal data setting, Y need not be scalar or even finite-dimensional; however, we now make the associated stochastic process implicit and instead begin with a specific filtration $(\mathcal{Y}_t : t \in T)$ of $\mathcal{Y}$ that, as t ranges over T, visits some or all of the possible sub-sigma algebras of $\mathcal{Y}$. This filtration supplies a partial ordering on T through the definition $s \leqslant t$ if and only if $\mathcal{Y}_s \subseteq \mathcal{Y}_t$, and specifies the various possible levels of coarsening with which we may gain information about $\mathcal{Y}$.

Heitjan & Rubin (1991, § 4.4) describes an example of coarsening, where children's ages are recorded to an unknown degree of precision. Ages may be rounded to the next lowest month, half year or full year, so that a child with a recorded age of 6 months may in fact be up to 11 months old. Let Y record the age of the child in months, rounded to the next lowest month. There are three possible sub-sigma algebras of $\mathcal{Y} = \sigma(Y)$, namely $\mathcal{Y}_{12} = \sigma(\{Y < 12\}, \{12 \leqslant Y < 24\}, \{24 \leqslant Y < 36\}, \ldots)$, $\mathcal{Y}_6 = \sigma(\{Y < 6\}, \{6 \leqslant Y < 12\}, \{12 \leqslant Y < 18\}, \ldots)$ and $\mathcal{Y}_1 = \mathcal{Y} = \sigma(\{Y = 1\}, \{Y = 2\}, \{Y = 3\}, \ldots)$, with $\mathcal{Y}_{12} \subseteq \mathcal{Y}_6 \subseteq \mathcal{Y}_1$. Here $T = \{12, 6, 1\}$, so a child with a recorded age of 18 months has an associated $\tau = 6$; while the age Y of the child could equal 18 exactly, all we know is that it lies in the set $\{18, \ldots, 23\}$. Suppose, for

argument's sake, that ages are in fact recorded to the next lowest month until one year of age, the next lowest half year until two years of age, and the next lowest full year thereafter. This constitutes a coarsening-at-random mechanism, because $P(\tau = t \mid \mathcal{Y}) = P(\tau = t \mid \mathcal{Y}_{12})$ for all t. Formulating coarsening in terms of sigma algebras neatly captures the spirit of set-valued variables introduced by Heitjan & Rubin (1991).

6. Discussion

Initially, our aim in this work was to provide a rigorous reinterpretation of the usual missingness-at-random formulation $P(M \mid Y) = P(M \mid Y_{\text{obs}})$ for those who, like ourselves, worry about such things. We hope that the version in Lemma 6 fits this bill. Seaman et al. (2013) point out that, interpreted literally, the symbol Y_{obs} might even tell us the value of M, but in fact no logical information about M is contained in any $\mathcal{Y}_m$, nor indeed in $\mathcal{Y}$ itself: for each nonempty set $A \in \mathcal{Y}$, the image $M(A)$ of A under M is simply the set $M(\Omega)$ of all possible values of M.

We believe that our work may have pedagogical value. Although we have attempted to convey our enthusiasm for the formalism of sigma algebras, an exactly equivalent version $P(M = m \mid Y) = P(M = m \mid Y_m)$ for all sets m does not rely on this concept. Those encountering this definition for the first time should see that there are many constituent subconditions, one for each possible subset $m \subseteq \{1, \ldots, n\}$, and that the conditioning object Y_m varies with m.

Our adaptedness requirement will seem natural to those familiar with stochastic processes, and provides further links between censoring and missing data (Aalen, 2007, 2012). The implied change of measure to a working independence setting also has a causal flavour: in causal inference, data-measurability is the key to identifiability of causal estimands, and employing stochastic bases $\{\Omega, \mathcal{F}, (\mathcal{F}_t)\}$ with causal interpretations seems to us a promising approach.

Acknowledgement

Odd Aalen, Daniel Commenges, Vern Farewell and Robin Henderson gave valuable advice during the writing of this paper. Rhian Daniel acknowledges support from a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society. Shaun Seaman was funded by the Medical Research Council.

Appendix

Examples of measure-theoretic quantities

Consider the simple setting introduced in § 3.1, in which a single binary Y_1 goes unobserved if $M = 0$, but is observed if $M = 1$. Recall that we defined Ω to be the four-point set $\{00, 01, 10, 11\}$ and that $\mathcal{F} = 2^\Omega$ was the power set of Ω. For a generic element $bc \in \Omega$, we let $Y_1(bc) = b$ and $M(bc) = c$.

In the case of finite sample spaces, we find it helpful to think of sigma algebras as partitioning the sample space into disjoint atoms, and have written the relevant sigma algebras to suggest this interpretation:

$$\begin{aligned} \mathcal{Y}_0 &= \sigma(\{00, 10, 01, 11\}), \\ \mathcal{M}_0 = \mathcal{F}_0 = \mathcal{M}_1 = \mathcal{M} &= \sigma(\{00, 10\}, \{01, 11\}), \\ \mathcal{Y}_1 = \mathcal{Y} &= \sigma(\{00, 01\}, \{10, 11\}), \\ \mathcal{D} = \mathcal{F}_M &= \sigma(\{00, 10\}, \{01\}, \{11\}), \\ \mathcal{F}_1 = \mathcal{F} &= \sigma(\{00\}, \{01\}, \{10\}, \{11\}). \end{aligned}$$

For any such sigma algebra $\mathcal{G} \subseteq \mathcal{F}$, we may uniquely associate one of its corresponding atoms with each point in the sample space, so that $A_{\mathcal{G}}(\omega) \in \mathcal{G}$ is the atom containing ω. For example, $A_{\mathcal{Y}}(00) = \{00, 01\}$ and $A_{\mathcal{D}}(00) = \{00, 10\}$. Strictly speaking, these are only atoms under measures that assign them positive probability; we shall implicitly assume this to be the case for the measures that we go on to describe. For any such measures P and Q, we claim that

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{G}}(\omega) = \frac{P\{A_{\mathcal{G}}(\omega)\}}{Q\{A_{\mathcal{G}}(\omega)\}}$$

for all ω, as might reasonably be expected of a quantity we describe as a likelihood ratio. This can be verified by applying the measure-theoretic definition of the conditional expectation $Q(\mathrm{d}P/\mathrm{d}Q \mid \mathcal{G})$ to the constituent atoms $A_{\mathcal{G}}$ of each set $A \in \mathcal{G}$.

Let us now be more specific about the probability measures in question. We let $P(Y_1 = 1) = p$ and $P(M = 1 \mid Y_1 = y) = p_y$, say, with similar notation for Q, so that

$$P(\{bc\}) = p^b(1-p)^{1-b}p_b^c(1-p_b)^{1-c}, \quad Q(\{bc\}) = q^b(1-q)^{1-b}q_b^c(1-q_b)^{1-c}$$

for a generic element $bc \in \Omega$. This notation allows us to emphasize that it is really the relative success of p and q in explaining the distribution of Y_1 that is assumed to be of principal scientific interest. We have recycled some notation here; p_b and q_b are deterministic and distinct from the stochastic processes (p_m) and (q_m) used in the main body of the paper. It is now straightforward to write down $\mathrm{d}P/\mathrm{d}Q$:

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) = \begin{cases} (1-p)(1-p_0)/(1-q)(1-q_0), & \omega = 00, \\ (1-p)p_0/(1-q)q_0, & \omega = 01, \\ p(1-p_1)/q(1-q_1), & \omega = 10, \\ pp_1/qq_1, & \omega = 11. \end{cases}$$

To evaluate its restriction to $\mathcal{D}$, we take each of the three atoms $A_{\mathcal{D}}$ of $\mathcal{D}$ in turn and get

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{D}}(\omega) = \begin{cases} \{(1-p)(1-p_0) + p(1-p_1)\}/\{(1-q)(1-q_0) + q(1-q_1)\}, & \omega \in \{00, 10\}, \\ (1-p)p_0/(1-q)q_0, & \omega = 01, \\ pp_1/qq_1, & \omega = 11. \end{cases}$$

Even more simply,

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{Y}}(\omega) = \begin{cases} (1-p)/(1-q), & \omega \in \{00, 01\}, \\ p/q, & \omega \in \{10, 11\}, \end{cases}$$

and

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{M}}(\omega) = \begin{cases} \{(1-p)(1-p_0) + p(1-p_1)\}/\{(1-q)(1-q_0) + q(1-q_1)\}, & \omega \in \{00, 10\}, \\ \{(1-p)p_0 + pp_1\}/\{(1-q)q_0 + qq_1\}, & \omega \in \{01, 11\}. \end{cases}$$

For a generic element $bc \in \Omega$, the measures P' and Q' of § 3.2 have corresponding probabilities

$$P'(\{bc\}) = p^b(1-p)^{1-b}\left(\frac{1}{2}\right)^c\left(1-\frac{1}{2}\right)^{1-c}, \quad Q'(\{bc\}) = q^b(1-q)^{1-b}\left(\frac{1}{2}\right)^c\left(1-\frac{1}{2}\right)^{1-c}$$

so that $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{Y}} = (\mathrm{d}Q/\mathrm{d}Q')|_{\mathcal{Y}} = 1$ as required, but now $\mathcal{Y}$ and $\mathcal{M}$ are independent under P' and Q'; the latter represent our working independence assumption. Replacing p_b and q_b by $1/2$ is an arbitrary

choice; any two constant, positive probabilities will do. By inspecting the values taken by $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}|\mathcal{M}}$ on its three atoms, or alternatively by applying Lemma 4, we see that

$$\left.\frac{\mathrm{d}P'}{\mathrm{d}Q'}\right|_{\mathcal{D}|\mathcal{M}}(\omega) = Q'\left(\left.\frac{\mathrm{d}P'}{\mathrm{d}Q'}\right|_{\mathcal{Y}}\right)(\omega) = \begin{cases} 1, & \omega \in \{00, 10\}, \\ (1-p)/(1-q), & \omega = 01, \\ p/q, & \omega = 11 \end{cases}$$

is our working independence conditional likelihood ratio, while

$$\left.\frac{\mathrm{d}P}{\mathrm{d}Q}\right|_{\mathcal{F}|\mathcal{Y}}(\omega) = \begin{cases} (1-p_0)/(1-q_0), & \omega = 00, \\ p_0/q_0, & \omega = 01, \\ (1-p_1)/(1-q_1), & \omega = 10, \\ p_1/q_1, & \omega = 11 \end{cases}$$

specifies the conditional likelihood ratio associated with the so-called missingness mechanism.

For the factorization of Theorem 1 to hold, we see that we require $\{(1-p)(1-p_0)+p(1-p_1)\}/\{(1-q)(1-q_0)+q(1-q_1)\}$ simultaneously to equal both $(1-p_0)/(1-q_0)$ and $(1-p_1)/(1-q_1)$, because $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{D}}$ is constant on $\{00, 10\}$ while $(\mathrm{d}P/\mathrm{d}Q)|_{\mathcal{F}|\mathcal{Y}}$ need not be. There are two cases in which both equalities hold: either $p_0 = p_1$ and $q_0 = q_1$, or $p = q$ and $(1-p_0)(1-q_1) = (1-p_1)(1-q_0)$. In the former case, $\mathcal{Y}$ and $\mathcal{M}$ are independent, and hence the missing-at-random assumption is trivially satisfied. The latter case is uninteresting since we presumably set out to compare distinct p and q; nevertheless, it shows that missingness at random is not a necessary condition for the factorization to hold.

Proof of Lemma 5

Formally, M induces the conditional measure μ defined by $\mu(D, \omega) = P(M^{-1}(D) \mid \mathcal{Y})(\omega)$ for any $D \subseteq \{1, \ldots, n\}$ and $\omega \in \Omega$, with a similar definition for μ' in terms of P'. We label the conditional densities of μ and μ', taken with respect to counting measure ν, as $p_m = (\mathrm{d}\mu/\mathrm{d}\nu)(m, \cdot) = P(M = m \mid \mathcal{Y})$ and p'_m, defined equivalently. From their definitions, p_m and p'_m are random variables, and consequently both (p_m) and (p'_m) may be viewed as stochastic processes indexed by m. We now prove that p_M/p'_M, the ratio of these density processes evaluated at the random stopping time M, is indeed the desired conditional likelihood ratio $(\mathrm{d}P/\mathrm{d}P')|_{\mathcal{F}|\mathcal{Y}}$. We do so by showing that for any set $A \in \mathcal{F}$, p_M/p'_M converts the conditional probability of A under P', given $\mathcal{Y}$, to the corresponding conditional probability under P.

Let $A \in \mathcal{F}$, and recall that any such A equals BC for some $B \in \mathcal{Y}$ and $C \in \mathcal{M}$, so that $P'(A \times p_M/p'_M \mid \mathcal{Y}) = B \times P'(C \times p_M/p'_M \mid \mathcal{Y})$. But by definition of $\mathcal{M}$ we may write $C = M^{-1}(D) = D_M$, say, for some $D \subseteq \{1, \ldots, n\}$. A change of variables then allows us to write $P'(D_M \times p_M/p'_M \mid \mathcal{Y}) = \mu'(D \times p/p', \cdot)$. Now $\mathrm{d}\mu/\mathrm{d}\mu' = p/p'$, whence we have $P'(A \times p_M/p'_M \mid \mathcal{Y}) = B \times \mu(D, \cdot) = B \times P(C \mid \mathcal{Y}) = P(A \mid \mathcal{Y})$ as required.

REFERENCES

AALEN, O. O. (2007). Contribution to the discussion of 'Longitudinal data with dropout: Objectives, assumptions and a proposal' by P. J. Diggle, D. Farewell and R. Henderson. *Appl. Statist.* **56**, 538–9.

AALEN, O. O. (2012). Armitage lecture 2010: Understanding treatment effects: The value of integrating longitudinal data and survival analysis. *Statist. Med.* **31**, 1903–17.

ANDERSEN, P. K., BORGAN, O., GILL, R. D. & KEIDING, N. (1996). *Statistical Models Based on Counting Processes*. New York: Springer.

CHANG, J. T. & POLLARD, D. (1997). Conditioning as disintegration. *Statist. Neer.* **51**, 287–317.

COMMENGES, D. & GEGOUT-PETIT, A. (2015). Likelihood inference for incompletely observed stochastic processes: Ignorability conditions. *arXiv:* math/0507151v2.

DIGGLE, P., HEAGERTY, P., LIANG, K.-Y. & ZEGER, S. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

Doretti, M., Geneletti, S. & Stanghellini, E. (2017). Missing data: A unified taxonomy guided by conditional independence. *Int. Statist. Rev.* **86**, 189–204.
Farewell, D. M., Huang, C. & Didelez, V. (2017). Ignorability for general longitudinal data. *Biometrika* **104**, 317–26.
Gill, R. D., van der Laan, M. J. & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proc. First Seattle Sympos. Biostatistics*, D. Y. Lin & T. R. Fleming, eds., Lecture Notes in Statistics. New York: Springer, pp. 255–94.
Grüger, J., Kay, R. & Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics* **47**, 595–605.
Hedeker, D. & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychol. Meth.* **2**, 64–78.
Heitjan, D. F. & Rubin, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19**, 2244–53.
Hoffman-Jørgensen, J. (1994). *Probability With a View Towards Statistics*, vol. II. Boca Raton, Florida: CRC Press.
Jacobsen, M. & Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23**, 774–86.
Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W. & Wang, J.-L., eds. (2014). *Past, Present, and Future of Statistical Science*. Boca Raton, Florida: Chapman and Hall/CRC.
Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Assoc.* **90**, 1112–21.
Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken, New Jersey: Wiley.
Lu, G. & Copas, J. B. (2004). Missing at random, likelihood ignorability and model completeness. *Ann. Statist.* **32**, 754–65.
Mealli, F. & Rubin, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102**, 995–1000.
Molchanov, I. (2006). *Theory of Random Sets*. New York: Springer.
Molenberghs, G., Beunckens, C., Sotto, C. & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Statist. Soc.* B **70**, 371–88.
Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.
Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge: Cambridge University Press.
Robins, J. M. & Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statist. Med.* **16**, 39–56.
Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. London: Routledge.
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–92.
Seaman, S., Galati, J., Jackson, D. & Carlin, J. (2013). What is meant by 'missing at random'? *Statist. Sci.* **28**, 257–68.
Sweeting, M. J., Farewell, V. T. & Angelis, D. D. (2010). Multi-state Markov models for disease progression in the presence of informative examination times: An application to hepatitis C. *Statist. Med.* **29**, 1161–74.

[*Received on* 7 *December* 2018. *Editorial decision on* 23 *December* 2020]