On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data

Peer-reviewed author version

MOLENBERGHS, Geert; Kenward, Michael G.; AERTS, Marc; VERBEKE, Geert; Davidian, Marie; Rizopoulos, Dimitris & Tsiatis, Anastasios A. (2014) On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: Sequential trials, random sample sizes, and missing data. In: Statistical methods in medical research, 23 (1), p. 11-41.

DOI: 10.1177/0962280212445801
Handle: http://hdl.handle.net/1942/14671

On Random Sample Size, Ignorability, Ancillarity, Completeness, Separability, and Degeneracy: Sequential Trials, Random Sample Sizes, and Missing Data

Geert Molenberghs[1,2] **Michael G. Kenward**[3] **Marc Aerts**[1]

Geert Verbeke[2,1] **Anastasios A. Tsiatis**[4] **Marie Davidian**[4]

Dimitris Rizopoulos[5]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[3] *Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London WC1E7HT, United Kingdom*

[4] *Department of Statistics, North Carolina State University, Raleigh, NC, U.S.A.*

[5] *Department of Biostatistics, Erasmus University Medical Center, NL-3000 CA Rotterdam, the Netherlands*

Abstract

The vast majority of settings for which frequentist statistical properties are derived assume a fixed, *a priori* known sample size. Familiar properties then follow, such as, for example, the consistency, asymptotic normality, and efficiency of the sample average for the mean parameter, under a wide range of conditions. We are concerned here with the alternative situation in which the sample size is itself a random variable which may depend on the data being collected. Further the rule governing this may be deterministic or probabilistic. There are many important practical examples of such settings, including missing data, sequential trials, and informative cluster size. It is well known that special issues can arise when evaluating the properties of statistical procedures under such sampling schemes, and much has been written about specific areas[3–4]. Our aim is to place these various related examples into a single framework derived from the joint modeling of the outcomes and sampling process, and so derive generic results that in turn provide insight, and in some cases practical consequences, for different settings. It is shown that, even in the simplest case of estimating a mean, some of the results appear counter-intuitive. In many examples the sample average may exhibit small sample bias and, even when it is unbiased, may not be optimal. Indeed there may be no minimum variance unbiased estimator for the mean. Such results follow directly from key attributes such as non-ancilliarity of the sample size, and incompleteness of the minimal sufficient statistic of the sample size and sample sum. Although our results have direct and obvious implications for estimation following group sequential trials, there are also ramifications for a range of other settings, such as random cluster sizes, censored time-to-event data, and the joint modeling of longitudinal and time-to-event data. Here we use the simplest sequential group sequential setting to develop and explicate the main results. Some implications for random sample sizes and missing data are also considered. Consequences for other related settings will be considered elsewhere.

Some Keywords: Frequentist Inference; Generalized Sample Average; Informative Cluster Size; Joint Modeling; Likelihood Inference; Missing at Random; Random Cluster Size.

1 Introduction

In much conventional statistical methodology, it assumed that the sample has a fixed and *a priori* known size. There exist many practical settings however in which sample sizes are, by contrast, random, and the derivation of the frequentist properties of statistical procedures in such cases raises fundamental issues that require careful attention. There are several distinct classes of problem of this type. First, data may be missing in the sense that fewer observations are collected than planned: the sample of units may be smaller than envisaged, or a set of observations from a unit, multivariate or longitudinal, may be incomplete[1]. Second, in (group) sequential trials, the final sample size depends on the point at which accumulating evidence crosses a pre-specified boundary[2]. Third, the sample size can be completely random (CRSS), in the sense that it does not depend on the observations, either collected already or still to be collected[3–4]. Here we incorporate such settings into a general framework for problems with non-fixed sampling schemes. To allow us to develop the main ideas in a simple way, but which still allows explication of the key problematic issues, we focus on the generic case where a sample size N can take only one of two values: n or $2n$. Having developed a full theory for this setting we then consider in turn the three particular settings mentioned above. A large literature already exists on various classes of the current problem. One of most familiar, and well worked concerns group sequential trials. Several authors have pointed to fundamental problems with the frequentist perspective of conventional estimators, when applied after conducting a sequential trial[5–8], whether they take the form of simple sample averages or those obtained through maximum likelihood. By incorporating a probit model for the stopping rule into the simple two sample size generic case described above, combined with normally distributed outcomes, we are able to develop a sufficiently rich setting for development, that includes the simple sequential sampling scheme as a limiting special case. By applying the work of Liu and Hall[8] in this setup, we are able to uncover the non-trivial implications of a random sample size, essentially making use of two important properties: (a) excluding the CRSS case, the sample size is non-ancillary given the sum of the observations made regardless of whether the stopping rule is probabilistic (as in many missing data applications) or deterministic (as in sequential trials); (b) the pair made up of the accumulating sample sum and the sample size is a minimal sufficient statistic that is not complete.

Using the above we show that the classical sample average is generally biased, but asymptotically unbiased. An unbiased estimator can be obtained from the conditional likelihood, where the conditioning is on the (non-ancillary) sample size. One consequence of this lack of ancillarity is that,

unexpectedly, the conditional estimator has larger mean squared error than the simple sample average, a result which follows from the joint likelihood. This underscores the point that, in contrast to some claims made in the literature, such an average is a perfectly legitimate estimator. The finite-sample bias in the sample average is zero only in the CRSS case. Even then, it is not unique in that a whole class of so-called generalized sample average estimators can be defined, all of which are unbiased. This enables us to show that the ordinary sample average is only asymptotically optimal. Indeed there is no uniformly optimal unbiased estimator in finite samples.

In addition to the three settings which we study in detail in this paper, there are several additional classes of problem that fit within our generic framework. We now list four of these, continuing the list of the second paragraph in this section. Fourth, so-called 'informative cluster sizes', in which the cluster sizes may themselves depend on the outcomes of interest. Fifth, censored time-to-event data have much in common with missing data. Sixth, the joint modeling of longitudinal outcomes and time-to-event data is an extension of various earlier settings, including incomplete longitudinal data, censored time-to-event data, and informative cluster sizes. Seventh, a longitudinal outcome may be paired with a random measurement occasion schedule. The general results derived in this paper also apply to these settings. The main results developed here will be applied to these four settings elsewhere.

This paper is organized as follows. In Section 2, a general framework is sketched, encompassing the various settings mentioned earlier. Key technical and modeling concepts are introduced. The main ideas are developed in context of a sequential trial with probabilistic stopping rule; it is the focus of Section 3. Ramifications for random sample sizes are discussed in Section 4, while in Section 5 we focus on missing data. Some technical details are deferred to appendices.

2 Generic Setting as 'Joint Modeling'

Consider the generic model for a pair of random variables $\boldsymbol{Y}_i$ and $\boldsymbol{C}_i$:

$$\begin{aligned} f(\boldsymbol{y}_i, \boldsymbol{c}_i|\boldsymbol{\theta}, \boldsymbol{\psi}) &= f(\boldsymbol{y}_i|\boldsymbol{\theta}) \cdot f(\boldsymbol{c}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) \quad (1) \\ &= f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{\theta}, \boldsymbol{\psi}) \cdot f(\boldsymbol{c}_i|\boldsymbol{\theta}, \boldsymbol{\psi}), \quad (2) \end{aligned}$$

where $\boldsymbol{Y}_i$ are outcomes of interest and $\boldsymbol{C}_i$ is a context-specific augmentation, such as sample size or missing-data pattern. Here $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are unknown fixed parameter vectors. In line with work by Rubin

and Little[1–9–11], (1) is the selection model factorization and (2) is the pattern-mixture factorization.

Also, following common practice in the missing-data literature, the mechanism $f(\boldsymbol{c}_i|\boldsymbol{y}_i, \boldsymbol{\psi})$ can be classified according to the following taxonomy. Missing data are *missing complete at random* (MCAR), if $f(\boldsymbol{c}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{c}_i|\boldsymbol{\psi})$, and *missing at random* (MAR), if $f(\boldsymbol{c}_i|\boldsymbol{y}_i, \boldsymbol{\psi}) = f(\boldsymbol{c}_i|\boldsymbol{y}_i^o, \boldsymbol{\psi})$, where BY_i^o is the observed portion of the sequence. Otherwise, the mechanism is *missing not at random* (MNAR). We can make a distinction between Rubin's original definition of these and a frequentist definition according to whether these particular forms are assumed to hold just for the data observed, or for all possible data sets respectively. Note that these concepts are not restricted to the missing data case, but rather apply to all of the above settings.

The seven instances of the previous section can be formalized as follows.

1. **Sequential trials.** Consider a simple sequential trial, where n measurements Y_i are observed, after which a stopping rule is applied and, depending on the outcome, another set of n measurements is or is not taken. Then, $\boldsymbol{Y}$ is the $(2n \times 1)$ vector of outcomes that could be collected and $\boldsymbol{C} = N$ is the realized sample size, that is, $N = n$ or $N = 2n$. This setting was considered by[12]. More generally, a sequential trial would allow for size $N = n_i < \cdots < n_L$, for pre-specified L.

2. **Missing longitudinal data.** Here $\boldsymbol{Y}_i$ is an sequence of planned measurements of length n_i for subject i in a trial of size N, and $\boldsymbol{C}_i$ is a vector of indicators, with $C_{ij} = 1$ if measurement Y_{ij} is taken and 0 otherwise. It is customary to decompose $\boldsymbol{Y}_i = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$, the observed and missing components, respectively.

3. **Completely random sample size.** Here $\boldsymbol{Y}$ is an outcome vector, as above, but with $C = N$ a CRSS, taking values $N = 0, \ldots, m$, with probabilities that are independent of $\boldsymbol{Y}$.

4. **Clusters of random size.** Assume that data are sampled in the form of clusters. Then $\boldsymbol{Y}_i$ is the set of outcomes for cluster $i = 1, \ldots, N$ and $\boldsymbol{C}_i = t_i$ is the observed cluster size. Two cases can be distinguished. First, some members of the cluster may fail to provide a measurement, implying that $t_i \leq n_i$, with n_i the number of cluster members. Second, in the case that there are no missing data, that is $t_i \equiv n_i$, it is still possible that the cluster size contain relevant information about the parameter of interest.

5. **Censored time-to-event data.** Now, $\boldsymbol{Y}_i = T_i$, a time-to-event for subject $i = 1, \ldots, N$ and

$\boldsymbol{C}_i = C_i$ is a censoring time.

6. **Joint models for longitudinal and time-to-event data.** In this setting, $\boldsymbol{Y}_i$ is a longitudinal outcome, as in the second setting, and $\boldsymbol{C}_i = (T_i, C_i)$ is the pair of event and censoring times, as in the previous setting.

7. **Random measurement times.** A situation closely related to the previous ones is that of random measurement times. By jointly considering the outcomes $\boldsymbol{Y}_i$ and the measurement times $\boldsymbol{T}_i$, this relates directly to the previous setting, as well as to the missing data and random cluster size settings.

In some of the settings above, covariates, X_i say, may be present. Without loss of generality these will be suppressed from notation, throughout.

A fundamental concept, also due to Rubin[9], is that of *ignorability*. For pure likelihood or Bayesian inferences, and assuming MAR, it is well known that inferences about $\boldsymbol{\theta}$ can proceed by merely using $f(\boldsymbol{y}_i^o|\boldsymbol{\theta})$, i.e., without explicitly modeling the missing-data mechanism. This is, provided the regularity condition of *separability* holds. (Note that this concept can be transported to generic setting (1), by letting $\boldsymbol{Y}_i^o$ be the observed portion in the first factor on the right hand side.) Formally, this implies that the parameter space of $(\boldsymbol{\theta}', \boldsymbol{\psi}')'$ is equal to the Cartesian product of their individual product spaces. Informally stated, this means that the missing-data model does not contain information about the outcome model parameter. Even when separability is not satisfied, the consequence will be efficiency loss, but validity of the estimation method will not be altered. In other words, $\boldsymbol{C}_i$ can be considered ancillary in the sense of Cox and Hinkley[13](pp. 32–35). We will see that this is not true for all situations. These considerations imply that ignorability may fail to hold for four reasons. First and most directly, we may be in the likelihood or Bayesian framework but with an MNAR mechanism operating. Second, ignorability does not hold in the likelihood and Bayesian framework, under MAR but in a non-separable situation; as stated above, this is not an important issue in practice. Third, under frequentist inferences, both MAR and MNAR are generally non-ignorable. Fourth, for pure likelihood and Bayesian inferences, with MAR and separability holding, ignorability in the selection model decomposition (1) does not translate to the pattern-mixture model (2), as is clear from the presence of both $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ in both factors of (2). Of course, one can consider ignorability in the pattern-mixture context, by directly parameterizing the PMM model; in this case, ignorability does not carry over to the selection model setting. We learn from this that the ignorability

is parameterization specific. When ranging over the seven settings described above, all of these situations may occur.

There may be instances where it is natural to model $f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{\theta}^*)$ on the one hand, but scientific interest lies with $f(\boldsymbol{y}_i|\boldsymbol{\theta}^*, \boldsymbol{\psi}^*)$ on the other. The asterisk refer to the fact that these are pattern-mixture parameters from (2) which are then transported to (1). This could be the case for example for the clustered-data setting. Then, the connection between both decompositions implies a natural *weight*:

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}^*, \boldsymbol{\psi}^*) = f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{\theta}^*) \cdot \frac{f(\boldsymbol{c}_i|\boldsymbol{\psi}^*)}{f(\boldsymbol{c}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*)} = f(\boldsymbol{y}_i|\boldsymbol{c}_i, \boldsymbol{\theta}^*) \cdot w_i(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*). \tag{3}$$

Here, 'natural weight' is used in the sense of being implied by a joint model, rather than following from *ad hoc* specification. Note that the above fourth case for non-ignorability now reverses: while the pattern-mixture formulation may satisfy ignorability, the derived selection model is unlikely to. The above is not speficic to the missing-data context but applies to all cases considered.

It is useful to point out the connection between ignorability and *ancillarity* in the sense of Cox and Hinkley[13]. These authors define an ancillary statistic T as one that complements a minimally sufficient statistic S such that, given S, T does not contain information about the parameter of interest. For example, when estimating the mean of a normally distributed population with mean μ and variance 1, the sample size $T = n$ is ancillary given the sample sum or sample average. This is certainly true for a sample size fixed by design and also for a random sample provided the law governing the sample size does not depend on μ. The general sequential trial case considered in this paper is a clear counterexample to this.

A further property that has an important role to play in the present context is that of *completeness*[14](pp. 285–286). A statistic $s(Y)$ of a random variable Y, with Y belonging to a family P_θ, is complete if, for every measurable function $g(\cdot)$, $E[g\{s(Y)\}] = 0$ for all θ, implies that $P_\theta[g\{s(Y)\} = 0] = 1$ for all θ. The relevance of completeness in the present setting arises in two ways. First, from the the Lehman-Scheffé theorem[14], if a statistic is unbiased, complete, and sufficient for some parameter θ, then it is the best mean-unbiased estimator for θ. This result is important in framing the problems that occur with estimation following group sequential trials, a point to which we return in Section 3. Second, the connection with ancillarity follows from Basu's theorem[14–15] (p. 287): a statistic that is both complete and sufficient is independent of any ancillary

statistic.

The above indicates that lack of completeness might correspond to violation of ignorability. Noting that these theorems are implications rather than equivalences, we will see that there are plenty of counterexamples in the sequential-trial setting. This means that a conventional statistical analysis, in line with what would be undertaken with complete data and based only upon $f(\boldsymbol{y}_i|\boldsymbol{\theta})$ in (1), may lead to bias. The lack of completeness may imply that there is no uniformly best way to overcome this problem, a point that is taken up in the next section.

We now consider the 'weights' in (3). These provide a natural connection with inverse probability weights; i.e., they provide a way to re-compose the population of interest from the sample data, although we do need to make the formal distinction between this type of of weighting which applies to the likelihood itself and the more familiar inverse probability or Horvitz-Thompson type[16] which essentially applies to the scores or estimating equations, and strictly lies outside the likelihood framework. By definition, the 'weights' can be removed in the ignorable cases, but are needed in all others. We will consider various cases in sections to follow. Evidently, the connection between both frameworks is problematic if the numerator $f(\boldsymbol{c}_i|\boldsymbol{y}_i, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*)$ of the 'weights' $w_i \equiv w_i(\boldsymbol{y}_i, \boldsymbol{c}_i, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*)$ is zero over a subset of the sample space or, more precisely, are not bounded away from zero. We will refer to these problematic situations as *degenerate* or *deterministic*, each of the terms stressing different aspects of the problem. This will occur for some incomplete-data and sequential-trial settings. In contrast to the semi-parametric case, this issue does not occur in the parametric situation. Nevertheless, caution is needed: the problem has been transferred to one of extrapolation, that is, unverifiable distributional assumptions have to be made about the model governing the sample space that cannot be reached.

A final remark is that in some cases there may fail to be replication to ensure estimability of all parameters, as we will see in the sequential-trial case. This is why the parameters governing the stopping rule, whether deterministic or probabilistic, will be assumed known. This does not imply a limitation.

3 Sampling With a Deterministic or Probabilistic Stopping Rule

Sequential trials have a long history[17], and throughout this, estimation following (group) sequential trials has been both of interest and controversial. This is true for point estimation[5–8] as well as

estimation of precision[12–18–20]. Much of the relatively early work has been summarized by Whitehead and colleagues[21].[20] point out that although maximum likelihood estimators are not affected by stopping rules, their sampling distribution is. This follows from ignorability, while the likelihood itself is unaffected by the stopping rule, frequentist aspects based on sampling distributions are. It was precisely this issue that was studied in Kenward and Molenberghs[12]. The same issue emerges in a different form in Hughes and Pocock[6], who observe that there is usually bias towards overestimation of the treatment effect; they aim at reducing this bias using prior information in a Bayesian analysis. Todd, Whitehead, and Facey[20] then provide methods for bias correction for the estimator. As will be seen see below, one has to be extremely careful with the concept of bias. The issue of bias will be revisited and it will be shown that many estimators, including the sample average, are asymptotically unbiased. Tsiatis, Rosner, and colleagues[18] and[19] study precision estimation after group sequential trials in terms of the joint sampling distribution of a test statistic K, obtained after T looks on the one hand, and T, the number of looks, on the other. Thus, these authors also considered joint modeling in the broad sense of Section 2. In particular, our simple sequential trial, introduced in Section 2 and further studied below, is a special case of this situation where T can take values 1 or 2, corresponding to $N = n$ and $N = 2n$, respectively. Emerson and Fleming[7] propose estimators within an ordering paradigm. Liu and Hall[8] cast the sequential trial test in the framework of a drift parameter θ of a Brownian motion $Y(t)$ stopped at time T, and show that the sufficient statistic $[T, Y(T)]$ is not complete for θ. Following our discussion of completeness in Section 2, it is not surprising that this means that there exist infinitely many unbiased estimators of θ, none of which has uniformly minimum variance.[8] then focus on the class of so-called truncation-adaptable unbiased estimators and find the minimum-variance member among them; thus, the problem is alleviated by making a restriction to smaller class of estimators than is usually the case. They also show that this estimator is identical to the one proposed by Emerson and Fleming[7].

The archetypical setting to be considered further was originally studied by Kenward and Molenberghs[12]. In particular, they suppose that n i.i.d. N$(\mu, 1)$ observations $Y_1, \ldots, Y_n$ are collected and, if the sample fails to satisfy a given stopping rule, a further n observations $Y_{n+1}, \ldots, Y_{2n}$ are collected, which have the same distribution. While n is not a random variable, the final sample size N is, taking possible values n or $2n$. Note that the setting is sufficiently general to be a paradigm for any (group) sequential trial, as will be made clear in Section 3.6. The inferential goal is to estimate μ, with an accompanying measure of precision. An obvious estimator is $\widehat{\mu} = N^{-1} \sum_{i=1}^{N} y_i$.

The aim of Kenward and Molenberghs[12] was to study the problems arising from such an approach, with emphasis on precision estimation. They argued that it is necessary to use the observed rather than the expected information matrix, because the setting is non-ignorable, in the sense that the corresponding likelihood is non-separable. Indeed, even though the stopping rule is conventionally defined in terms of the first n observations, and so independent of possibly missing outcomes, N does contain information about μ, given the observed outcomes. Equivalently, this means that N is not ancillary for μ given the observed data or, more precisely, a minimally sufficient statistic such as the average or sum of the observations. As is clear from the above literature review, the simple sample average is often surrounded with doubts about its sampling behavior. We will see below that it is a legitimate estimator, in a precisely defined sense. Somewhat surprisingly, we will see that this estimator is *not* constructed through conditioning on N, but rather is derived from from the joint likelihood for the data and the sample size. We will show that this is in complete agreement with the use of likelihood in an ignorable setting[9]. We will see, by contrast, that the conditional estimator takes a different form.

[12] proceeded by considering the full likelihood, using the pattern-mixture decomposition,

$$f(y_1, \ldots, y_N|N)f(N). \tag{4}$$

They focused on the limiting case where the trial is stopped if $\sum_{i=1}^{n} y_i < 0$, and continued otherwise. This rule is deterministic, in line with common practice. They showed that the marginal probability of stopping $\pi = \Phi(-\sqrt{n}\mu)$, with $\Phi(\cdot)$ the standard normal cumulative distribution function, and proceeded with the derivation of the corresponding log-likelihood and observed information. We will now frame the deterministic stopping rule as the limiting case of a probabilistic law, by considering the conditional probability of stopping:

$$\pi(N = n|\boldsymbol{y}_1) = \pi(\boldsymbol{y}_1) = F\left(\alpha + \frac{\beta}{n}k\right), \tag{5}$$

where

$$\boldsymbol{Y}_1 = (Y_1, \ldots, Y_n) \tag{6}$$

and $K = \sum_{i=1}^{n} Y_i$. Here, K is the sum over the observations available at the time of evaluation. So, at the time of the interim look, K is over n observations. K can also be evaluated after $2n$

observations for an experiment that continues after the interim look.

Even though the sample size N itself indicates whether or not stopping occurs, it is convenient to introduce a random variable Z that takes the value 1 if stopping occurs and 0 if not. Alternatively, the indicators $I(N = n)$ and $I(N = 2n)$ may be used. The full likelihood can be written:

$$L^* = \prod_{i=1}^{2n} \phi(y_i; \mu) \cdot F\left(\alpha + \frac{\beta}{n}k\right)^z \cdot \left\{1 - F\left(\alpha + \frac{\beta}{n}k\right)\right\}^{1-z}. \tag{7}$$

Here, $\phi(\cdot)$ is the standard normal density. Because Z is not replicated, α and β cannot be estimated from the data, so these parameters will be assumed fixed by design. This has no implications for the conceptual issues developed with this example. The observed data likelihood then takes the form:

$$L = \prod_{i=1}^{N} \phi(y_i; \mu) \cdot F\left(\alpha + \frac{\beta}{n}k\right)^z \cdot \left\{1 - F\left(\alpha + \frac{\beta}{n}k\right)\right\}^{1-z}. \tag{8}$$

Expressions (7) and (8) are identical, except for the sample size, which is $2n$ in the first case and N in the second. Because (7) starts from a selection model decomposition, it contains the marginal data model, rather than the one conditional on sample size. For the sample size, the reverse is true, in the sense that a conditional expression is given, rather than the marginal one. To reverse the factorization, as in (4), we proceed by writing

$$\boldsymbol{Y}_1 \sim N(\boldsymbol{\mu}, I_n), \tag{9}$$

where $\boldsymbol{\mu}$ is a vector of n copies of μ and I_n is the n-dimensional identity matrix. By taking $F(\cdot)$ in (5) to be the probit function $\Phi(\cdot)$, we can consider a latent stopping variable

$$T|\boldsymbol{y}_1 \sim N\left(\alpha + \frac{\beta}{n}k, 1\right), \tag{10}$$

underlying Z. From this, we can derive the joint distribution of $\boldsymbol{Y}_1$ and T, using standard multivariate normal manipulations:

$$\left(\begin{array}{c} \boldsymbol{Y}_1 \\ \hline T \end{array}\right) \sim N\left[\left(\begin{array}{c} \mu\mathbf{1}_n \\ \hline \alpha + \beta\mu \end{array}\right), \left(\begin{array}{c|c} I_n & \frac{\beta}{n}\mathbf{1}_n \\ \hline \frac{\beta}{n}\mathbf{1}_n' & 1 + \beta^2/n \end{array}\right)\right], \tag{11}$$

where $\mathbf{1}_n$ is an n-vector of ones. The marginal probability of stopping is therefore:

$$P(N = n) = P(Z = 1) = \Phi\left(\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}}\right). \tag{12}$$

Note that (12) depends on the parameter μ, implying that this pattern-mixture formulation non-separable. By contrast, although the observed data are present in the conditional stopping probability, μ is not, implying separability in the selection model formulation. In the next section, we will encounter an example of the reverse.

Continuing, the conditional probability for the outcomes is:

$$\begin{aligned} &f(y_1, \ldots, y_N | Z = z) = \\ &\frac{\prod_{i=1}^{N} \phi(y_i; \mu) \Phi\left(\alpha + \frac{\beta}{n}k\right)^z \left[1 - \Phi\left(\alpha + \frac{\beta}{n}k\right)\right]^{1-z}}{\Phi\left(\frac{\alpha+\beta\mu}{\sqrt{1+\beta^2/n}}\right)^z \left[1 - \Phi\left(\frac{\alpha+\beta\mu}{\sqrt{1+\beta^2/n}}\right)\right]^{1-z}}. \end{aligned} \tag{13}$$

The connection between the marginal outcome model (9) and its conditional counterpart (13) again draws attention to a natural 'weight' or connecting factor:

$$f(y_1, \ldots, y_N) = f(y_1, \ldots, y_N | Z = 1) \cdot w(y_1, \ldots, y_N, Z = 1)$$

with

$$w(y_1, \ldots, y_N, Z = 1) = \frac{\Phi\left(\frac{\alpha+\beta\mu}{\sqrt{1+\beta^2/n}}\right)}{\Phi\left(\alpha + \frac{\beta}{n}k\right)}. \tag{14}$$

A similar 'weight' applies to the case where $Z = 0$.

We now derived the limiting case in which the stopping rule is a deterministic function of the first n observations, i.e. we let $\beta \to \pm\infty$. Focussing on the positive limit, we obtain

$$\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}} \overset{\beta\to+\infty}{\longrightarrow} \sqrt{n}\mu.$$

This leads to the probability derived directly by[12], up to an immaterial sign reversal: we will choose, in what follows, to avoid the minus sign. In this limiting case, the marginal outcome model retains the form (9), but the three other expressions change. First, the conditional outcome model (13)

becomes

$$f(y_1, \ldots, y_N | Z = z) = \frac{\prod_{i=1}^{N} \phi(y_i; \mu)}{\Phi\left(\sqrt{n}\mu\right)^z \left[1 - \Phi\left(\sqrt{n}\mu\right)\right]^{1-z}}. \tag{15}$$

Second, (5) is given by $P(N = n|\boldsymbol{y}_i) = 1$ if $K > 0$ and 0 otherwise. Third, (12) takes the limiting form $P(N = n) = \Phi\left(\sqrt{n}\mu\right)$.

These results, in both probabilistic and limiting deterministic forms, have important implications which we now explore. First, when a joint modeling approach is used, beginning from (5) and (9), the kernel of the likelihood is just

$$L(\mu) \propto \prod_{i=1}^{N} \phi(y_i; \mu). \tag{16}$$

By contrast, for the approach that is conditional on $Z = z$, the more elaborate expression (13) applies. Hence each approach leads to different inferences. Whereas in Cox and Hinkley[13], in particular pages 31–35, it is argued that while conditioning on ancillary statistics is generally sensible, they also draw attention to the point that there is no universal method for constructing these. These authors also discuss a random sample size in this context. In their example, notwithstanding this randomness, ancillarity is maintained, in contrast to our example. We will see below that this lack of ancillarity leads to non-standard results.

Second, while the limiting deterministic case does not differ fundamentally from the general probabilistic one in a likelihood setting, we do see that the natural 'weight' (14) becomes

$$w(y_1, \ldots, y_N, Z = 1) = \frac{\Phi\left(\sqrt{n}\mu\right)}{I(K \geq 0)}. \tag{17}$$

with degeneracy in the denominator.

We now study the two likelihood estimators in more detail.

3.1 The Marginal and Conditional Likelihood Estimators

Beginning with the conditional model (13) we write, for convenience, $\widetilde{\alpha} = \alpha/\sqrt{1 + \beta^2/n}$ and $\widetilde{\beta} = \beta/\sqrt{1 + \beta^2/n}$. Further, let $\nu = \widetilde{\alpha} + \widetilde{\beta}\mu$. Evidently, $\nu = \nu(\mu)$ is a (simple linear) function of μ.

Consider first the case where $Z = 1$. The kernel of the likelihood $\ell(\mu)$, score function $S(\mu)$, and

Hessian $H(\mu)$ are:

$$\begin{aligned} \ell(\mu) &= -\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu)^2 - \ln\Phi(\nu), && (18)\\ S(\mu) &= \sum_{i=1}^{N}(y_i-\mu) - \widetilde{\beta}\cdot\frac{\phi(\nu)}{\Phi(\nu)}, && (19)\\ H(\mu) &= -N + \widetilde{\beta}^2\cdot[\nu\cdot\Phi(\nu)+\phi(\nu)]\cdot\frac{\phi(\nu)}{\Phi(\nu)^2}. && (20) \end{aligned}$$

When $Z=0$, the corresponding expressions are:

$$\begin{aligned} \ell(\mu) &= -\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu)^2 - \ln\left[1-\Phi(\nu)\right], && (21)\\ S(\mu) &= \sum_{i=1}^{N}(y_i-\mu) + \widetilde{\beta}\cdot\frac{\phi(\nu)}{1-\Phi(\nu)}, && (22)\\ H(\mu) &= -N - \widetilde{\beta}^2\cdot\{\nu\cdot[1-\Phi(\nu)]-\phi(\nu)\}\cdot\frac{\phi(\nu)}{[1-\Phi(\nu)]^2}. && (23) \end{aligned}$$

Evidently, for the joint approach, the estimator agrees with that for a simple sample from a normal distribution and we obtain the following familiar expressions:

$$\begin{aligned} \ell(\mu) &= -\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu)^2, && (24)\\ S(\mu) &= \sum_{i=1}^{N}(y_i-\mu), && (25)\\ H(\mu) &= -N. && (26) \end{aligned}$$

The simplicity of this estimator is a direct consequence of ignorability, given that the stopping rule satisfies MAR.

To illustrate the difference between the joint and conditional estimators, a brief simulation study has been conducted. A sample of size n was drawn from a $N(\mu,1)$ distribution. For given values of α and β, stopping rule (10) was applied. Whenever needed, a second sample, equally of size n was generated. For all cases, a simulation run consisted of 1000 samples. Results are summarized in Table 1. As indicated above, joint estimation is based on the simple sample average.

It is clear from (1) that both estimators are asymptotically unbiased, owing to their likelihood basis.

This will be studied further using the likelihood equations, in Section 3.4. Indeed, (25) is the marginal score, and its expectation should be taken over the entire space, where each $E(Y_i) = \mu$. For the other two, (19) and (22), expectation should be taken over the conditional space, given $Z = z$. Fortunately, the score itself is derived for this space, hence a consistent estimator follows based on standard likelihood theory. More formal arguments will be given in Section 3.4.

The results in Table 1 suggest that the precision is slightly higher for the joint approach rather than for the conditional one. This may appear counterintuitive at first sight, because in many practical applications a conditional approach is known to be more precise. However, this intuition draws from our common experience of conditioning on an *ancillary* statistic, which is not the case here. The superior precision for the joint estimator that is seen in Table 1 can in fact be predicted from the likelihood equations. The expected information in the joint approach is

$$I(\mu) = E(N) = n \cdot \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu) + 2n \cdot [1 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)] = n[2 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)], \tag{27}$$

where $\widetilde{\alpha}$ and $\widetilde{\beta}$ are as above. In the conditional case, this is

$$\begin{aligned} I_c(\mu) &= n - \widetilde{\beta}^2 \cdot \frac{\phi(\widetilde{\alpha} + \widetilde{\beta}\mu)}{\Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)^2} \left[(\widetilde{\alpha} + \widetilde{\beta}\mu) \cdot \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu) + \phi(\widetilde{\alpha} + \widetilde{\beta}\mu) \right] \\ &\quad + 2n + \widetilde{\beta}^2 \cdot \frac{\phi(\widetilde{\alpha} + \widetilde{\beta}\mu)}{[1 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)]^2} \left\{ (\widetilde{\alpha} + \widetilde{\beta}\mu) \cdot [1 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)] - \phi(\widetilde{\alpha} + \widetilde{\beta}\mu) \right\} \\ &= n[2 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)] - \frac{\widetilde{\beta}^2 \phi(\widetilde{\alpha} + \widetilde{\beta}\mu)^2}{\Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)[1 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)]} \end{aligned} \tag{28}$$

When $n \to \infty$, the information approaches

$$I_c(\mu) \overset{n \to +\infty}{\longrightarrow} n \left\{ [2 - \Phi(\alpha + \beta\mu)] - \frac{1}{n} \cdot \frac{\beta^2 \phi(\alpha + \beta\mu)^2}{\Phi(\alpha + \beta\mu)[1 - \Phi(\alpha + \beta\mu)]} \right\}.$$

It is then clear that the 'information deficit' will tend to zero when n tends to infinity.

In the limiting case, when $\widetilde{\beta} \to +\infty$, the second term in (28) approaches

$$\frac{n\phi(\sqrt{n}\mu)^2}{\Phi(\sqrt{n}\mu)[1 - \Phi(\sqrt{n}\mu)]}.$$

This term is non-zero for finite n but can be shown to approach 0 if $n \to \infty$.

We conclude that the conditional estimator is less precise than the joint one, in contrast to many

familiar settings such as contingency table analyses. The important feature here is conditioning is made on a non-ancillary statistic. We have also seen that the joint approach leads to the ordinary sample average, an estimator that has met with considerable concern in the past in the sequential setting. These and related issues are studied in Section 3.2.

As an aside, the problem that exists with use of the expected, as opposed to the observed, information matrix, as elucidated in Kenward and Molenberghs[12], remains. In summary, likelihood inference is valid, provided that the observed information matrix is used[3]. We return to the maximum likelihood estimators in Section 3.4.

3.2 Complete and Incomplete Sufficient Statistics

We now consider the role of completeness in this setting. For this, our argument owes much to that in Liu and Hall[8]. A sufficient statistic for a group sequential trial, of the type we are considering and in general, is (K, N) as defined above. Liu and Hall[8] show that (K, N) is incomplete in general group sequential trials. This means that the Lehmann-Scheffé theorem cannot be invoked (see Section 2); that is if a statistic is unbiased, complete, and sufficient for a parameter μ, then it is the best mean-unbiased estimator for μ. Of course, this in itself does not mean that there is no such optimal estimator, but Liu and Hall[8] show directly that an optimal estimator does not exist in the group sequential case. While our simple sequential trial has only two possible sample sizes n and $2n$, we have introduced a more general, probabilistic, stopping rule than is usually considered in the group sequential setting (5), and so we need to extend appropriately the results of Liu and Hall[8]. We demonstrate in this subsection that, given the lack of completeness of the sufficient statistics, there exist classes of unbiased estimators in this problem for which no member is uniformally optimal.

Continuing with the same example in which the outcomes are assumed to be normally distributed with mean μ and variance 1, and applying the results of Liu and Hall[8] to our sequential-trial case for either stopping at $N = n$ or continuing to $N = 2n$, we find that the joint density for K and N can be written

$$p_\mu(N, k) = p_0(N, k) \cdot \exp\left(k\mu - \frac{1}{2}n\mu^2\right)$$

with $p_0(N,k) = f_0(N,k)$ for k in the stopping region, i.e., $k < 0$, and 0 otherwise, for

$$\begin{aligned} f_0(n,k) &= \phi_n(k) \cdot I(k > 0), & (29) \\ f_0(2n,k) &= \int_{k=-\infty}^{k=0} f_0(z,n) \cdot \phi_n(k-z)\, dz, & (30) \end{aligned}$$

with $\phi_s(k)$ the normal density with mean 0 and variance s.

Now, when we apply the more general probabilistic stopping rule (5) with the probit form of $F(\cdot) = \Phi(\cdot)$, the deterministic case is obtained by letting $\beta \to +\infty$. Hence for, for general β, we obtain the corresponding joint densities by replacing the integration boundaries present in (29) and (30), by the corresponding probabilities:

$$\begin{aligned} p_0(n,k) &= \phi_n(k) \cdot \Phi\left(\alpha + \frac{\beta}{n}k\right), & (31) \\ p_0(2n,k) &= \int_{k=-\infty}^{k=+\infty} \phi_n(z) \cdot \left[1 - \Phi\left(\alpha + \frac{\beta}{n}k\right)\right] \cdot \phi_n(k-z)\, dz \\ &= \phi_{2n}(k) \cdot \left[1 - \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)\right], & (32) \end{aligned}$$

the derivation of which is straightforward but tedious. A sketch of the proof is given in Appendix A.

When $\beta \to +\infty$, (32) reduces to (30). On the other hand when $\beta = 0$, we recover the CRSS case. For this CRSS case, (31)–(32) reduces to:

$$\begin{aligned} p_0(n,k) &= \phi_n(k) \cdot \Phi, & (33) \\ p_0(2n,k) &= \phi_{2n}(k) \cdot (1 - \Phi), & (34) \end{aligned}$$

where $\Phi \equiv \Phi(\alpha)$, the random stopping probability. Note that this is a constant, indeed, because $\beta = 0$ and hence the dependence on the random variable K vanishes.

We now assume that a function $g(k,N)$ exists such that its expectation is zero. Such a function must satisfy, for all values of μ:

$$g(k,2n) \cdot p_0(2n,k) = -\int \phi_n(k-z) \cdot g(z,n) \cdot \phi_n(z) \cdot \Phi\left(\alpha + \frac{\beta}{n}z\right) dz. \quad (35)$$

Consider the following two examples.

Example 1. Choose

$$g(k,n) = \widetilde{\lambda}, \tag{36}$$

an arbitrary constant. Then it immediately follows from (35) that

$$g(k,2n) = -\widetilde{\lambda} \cdot \frac{\Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)}{1-\Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)}. \tag{37}$$

When $\beta = 0$, then the right hand side of (37) is constant and we can set $\widetilde{\lambda} = \lambda(1-\Phi)$, leading to $g(k,n) = \lambda(1-\Phi)$ and $g(k,2n) = -\lambda\Phi$.

Example 2. Choose

$$g(k,n) = \frac{\lambda}{\Phi\left(\alpha + \frac{\beta}{n}k\right)}, \tag{38}$$

with λ a given constant, then

$$g(k,2n) = -\frac{\lambda}{1-\Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)}. \tag{39}$$

Such $g(k,N)$ functions lead to entire classes of estimators. To see this, assume that an estimator for μ is available, $\widehat{\mu}$, say. For example, $\widehat{\mu}$ could be the sample average

$$\widehat{\mu} = \frac{1}{N}K \tag{40}$$

or a generalization of this:

$$\overline{\mu} = \frac{K}{N} \cdot [c \cdot I(N=n) + d \cdot I(N=2n)] = K \cdot \left[\frac{c \cdot I(N=n)}{n} + \frac{d \cdot I(N=2n)}{2n}\right], \tag{41}$$

for constant c and d. We will return to (40) and (41), but for now we concentrate on the construction of classes of functions.

Example 1 (continued). Choosing (36) and (37) for the special case of $\beta = 0$ leads to the following class of estimators:

$$\widehat{\mu}_\lambda = \overline{\mu} + \lambda \cdot [(1-\Phi)I(N=n) - \Phi I(N=2n)]\,. \tag{42}$$

It follows directly from the construction of $g(k,N)$ that $E(\overline{\mu}) = E(\widehat{\mu}_\lambda)$ and hence, if $\overline{\mu}$ is unbiased, then so is $\widehat{\mu}_\lambda$.

For the variance of (42), we obtain $\text{var}(\widehat{\mu}_\lambda) = \text{var}(\overline{\mu}) + \lambda^2\Phi(1-\Phi)$ which, within this class, is minimal for $\lambda = 0$. Hence, for $\beta = 0$, i.e., the CRSS case, the original estimator is more efficient than any member of the new class. This will change when $\beta \neq 0$. We also need to consider the basic estimator itself, e.g., either (40) or (41). before moving on to this, we first complete the second example.

Example 2 (continued). Choosing (38) and (39) produces the estimator

$$\widetilde{\mu}_\lambda = \overline{\mu} + \lambda \cdot \left[\frac{I(N=n)}{\Phi\left(\alpha + \frac{\beta}{n}k\right)} - \frac{I(n=2n)}{1-\Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)} \right]. \tag{43}$$

3.3 The Generalized Sample Average

The two examples above both take the form of augmented sample averages. It is, however, of interest to consider the sample average, or rather its generalization (41), in its own right. This is relevant here because several authors in the group sequential literature have reported that the the sample average is biased[6–20]. However, as shown at the beginning of Section 3.1, the sample average is the maximum likelihood estimator from the joint model.

Consider now the generalized sample average estimator (41). Clearly the arithmetic sample average is the special case of this for which for $c = d = 1$. Straightforward but tedious algebra leads to the expectation (Appendix A):

$$E(\overline{\mu}) = d\mu + (c-d)\mu\Phi(\nu) + \frac{2c-d}{2n}\frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu) \tag{44}$$

with $\nu = (\alpha + \beta\mu)/\sqrt{1 + \beta^2/n}$.

Consider first $\beta = 0$, then we have again that $\Phi \equiv \Phi(\alpha)$ and hence the estimator (41) is unbiased if and only if

$$d = \frac{1 - c\Phi}{1 - \Phi}. \tag{45}$$

Clearly, $c = d = 1$ is a solution, so the ordinary sample average is an unbiased estimator. An obvious question then concerns the optimal choices of c and d. From first principles (Appendix A), the variance can be derived as

$$\text{var}(\overline{\mu}) = \mu^2(c-1)^2\left(\frac{\Phi}{1-\Phi}\right) + \frac{1}{n}\Phi c^2 + \frac{1}{2n(1-\Phi)}(1 - c\Phi)^2, \tag{46}$$

which is minimal for

$$c_{\text{opt}} = 1 \; - \; \frac{1}{2n} \cdot \frac{1-\Phi}{\mu^2 + \frac{2-\Phi}{2n}}, \qquad d_{\text{opt}} = 1 \; + \; \frac{1}{2n} \cdot \frac{\Phi}{\mu^2 + \frac{2-\Phi}{2n}}. \tag{47}$$

These coefficients are different from 1 and are *not uniform* across μ. So, for finite samples the ordinary sample average is not optimal, but is asymptotically.

When $\beta \neq 0$, the expression (44) does not in general simplify. Suppose that we assume existence of a uniformly unbiased estimator, i.e., that there exist c and d such that (44) reduces to μ, for all μ, and in particular for $\mu = 0$. For this special case

$$0 = \frac{2c - d}{2n} \cdot \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \phi(\nu).$$

Given that $\beta \neq 0$, this expression leads to the condition $2c = d$. Substituting this into (44) produces

$$E(\overline{\mu}) = c\mu\left[2 - \Phi(\nu)\right],$$

which equals μ only if $c = [2 - \Phi(\nu)]^{-2}$. But given that $\Phi(\nu)$ is not constant but rather depends on μ, unless $\beta = 0$, we see that there can be no uniformly unbiased estimator for the generalized sample average type. In other words, a simple average estimator, that merely uses the observed measurements in a least-squares fashion, can never be unbiased unless $\beta = 0$.

It is insightful to study the generalized sample average's asymptotic bias. This has been done for

$\beta = 0$, with all unbiased solutions given by (45). All other choices for c and d will lead to bias that does not disappear asymptotically.

Turning to $\beta \neq 0$, we begin with the ordinary sample average, i.e. $c = d = 1$, which has expectation equal to

$$E(\widehat{\mu}) = \mu + \frac{1}{2n} \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \phi(\nu) \overset{n \to +\infty}{\longrightarrow} \mu + \frac{1}{2n} \beta \cdot \phi(\alpha + \beta\mu) \overset{n \to +\infty}{\longrightarrow} \mu. \tag{48}$$

In particular, when $\beta \to +\infty$, we see that

$$E(\widehat{\mu}) = \mu + \frac{1}{2\sqrt{n}} \cdot \phi(\sqrt{n}\mu) \overset{n \to +\infty}{\longrightarrow} \mu. \tag{49}$$

There exist other choices that also lead to asymptotically unbiased generalized sample averages. For $\beta \neq 0$ but finite, the expectation becomes

$$E(\overline{\mu}) \overset{n \to +\infty}{\longrightarrow} d\mu + (c - d)\mu\Phi(\alpha + \beta\mu), \tag{50}$$

which equals μ if and only if:

$$d = \frac{1 - c\Phi(\alpha + \beta\mu)}{1 - \Phi(\alpha + \beta\mu)}. \tag{51}$$

While (51) and (45) are similar, there is a crucial difference between these: the latter is independent of μ, while the former is not, except when $c = d = 1$. In other words, there is no uniformly asymptotically unbiased generalized sample average for finite, non-zero β, except for the ordinary sample average itself.

The situation is different for the deterministic $\beta \to \infty$, because then (50) becomes

$$E(\overline{\mu}) = d\mu + (c - d)\mu\Phi(\sqrt{n}\mu) + \frac{2c - d}{2\sqrt{n}}\phi(\sqrt{n}\mu) \overset{n \to +\infty}{\longrightarrow} \begin{cases} c\mu & \text{if } \mu > 0, \\ d\mu & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0. \end{cases} \tag{52}$$

This provides us with the interesting situation that, for positive μ, $c = 1$ yields an asymptotically unbiased estimator, regardless of d, with the reverse holding for negative μ. In the special case that $\mu = 0$, both coefficients are immaterial. In addition, we see here as well that the only uniform solution is obtained by requiring that the bias asymptotically vanishes for all values of μ, that is $c = d = 1$.

We have seen above that, even for $\beta = 0$, the sample average is not optimal, and that there is no uniform optimal solution, even though the sample average approximately is. However, the sample average is optimal in the restricted class of estimators that is invariant to future decisions. Indeed, if stopping occurs, then the choice of the coefficient c leads to an unbiased estimator, provided the appropriate d is chosen. However, this d will never be used as it pertains to 'future' observations. This can be avoided only by setting both coefficients to be equal, from which the conventional sample average emerges.

3.4 The Likelihood Estimators Revisited

We can also obtain the asymptotic unbiasedness of the sample average from the fact that it is the maximum likelihood estimator from the joint likelihood (16). In terms of the conditional likelihood, the estimator is obtained from the solution to the score equations, (19) and (22). These can be that can be reformulated as:

$$\widetilde{S}(\mu) = \frac{1}{N}\sum_{i=1}^{N} y_i - \mu - \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu) \cdot \left\{ \frac{I(N=n)}{n \cdot \Phi(\nu)} - \frac{I(N=2n)}{2n \cdot [1-\Phi(\nu)]} \right\}. \tag{53}$$

The expectation of (53) results from (48), combined with the observation that the probability of stopping is $\Phi(\nu)$:

$$\begin{aligned} E[\widetilde{S}(\mu)] &= \mu + \frac{1}{2n}\frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu) - \mu \\ &\quad - \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu) \cdot \left\{ \frac{1}{n \cdot \Phi(\nu)} \cdot \Phi(\nu) - \frac{1}{2n \cdot [1-\Phi(\nu)]} \cdot [1-\Phi(\nu)] \right\} = 0. \end{aligned}$$

Unbiasedness follows directly from the linearity of the score in the data. It follows that, while the conditional estimator is slightly less precise than the joint estimator, as follows from comparing (27) and (28), the former is nevertheless unbiased whereas the latter is only asymptotically so.

In other words, the difference between both score equations is bias-correcting. The correction is a non-linear function of μ and has no closed-form solution, underscoring the point that no simple algebraic function of K and N will lead to the same estimator.

Combining (27) with the bias, leads to the mean squared error for the joint estimator

$$\mathsf{MSE}(\widehat{\mu}) = \frac{1}{n[2-\Phi(\nu)]} + \frac{1}{4n^2}\widetilde{\beta}^2\phi(\nu)^2, \tag{54}$$

whereas (28) produces:

$$\mathsf{MSE}(\widehat{\mu}_c) \simeq \frac{1}{n[2-\Phi(\nu)]} + \frac{1}{[2-\Phi(\nu)]^2\Phi(\nu)[1-\Phi(\nu)]n^2}\widetilde{\beta}^2\phi(\nu)^2. \tag{55}$$

Comparing both expressions, we see that $G(\nu) = [2-\Phi(\nu)]^2\Phi(\nu)[1-\Phi(\nu)] < 4$, the inequality being strict. In fact, the maximal value for $G(\nu)$ equals 0.619. Hence, the joint estimator has the smallest MSE of both, even though the difference will be very small for moderate to large sample sizes. This holds regardless of the choice for α, β, and n, and of the true value of μ. For β finite and when $n \to \infty$, ν approaches $\alpha+\beta\mu$ and $\widetilde{\beta}$ approaches β. Then, $\Phi(\alpha+\beta\mu)$ and $\phi(\alpha+\beta\mu)$ become constant and the difference between the two expressions disappears because the second terms on the right hand sides of (54) and (55) are of the order of $1/n^2$.

A different argument applies in the deterministic case, when $\beta \to +\infty$, because then $\nu = \sqrt{n}\mu$ and $\widetilde{\beta} = \sqrt{n}$, producing

$$\begin{aligned}
\mathsf{MSE}(\widehat{\mu}|\beta\to\infty) &= \frac{1}{n[2-\Phi(\sqrt{n}\mu)]} + \frac{1}{4n}\phi(\sqrt{n}\mu)^2, \\
\mathsf{MSE}(\widehat{\mu}_c|\beta\to\infty) &\simeq \frac{1}{n[2-\Phi(\sqrt{n}\mu)]} + \frac{1}{[2-\Phi(\sqrt{n}\mu)]^2\Phi(\sqrt{n}\mu)[1-\Phi(\sqrt{n}\mu)]n}\phi(\sqrt{n}\mu)^2.
\end{aligned}$$

When $n \to \infty$, and using the same arguments as in Section 3.1, both expressions converge to $1/(2n)$ if the trial continues and $1/n$ if the trial stops, and the difference between them disappears.

In Section 3.1, we saw that the joint estimator appears to take a simpler form than the conditional one, but this is deceptive, because the latter in fact removes information, as is clear from (27) and (28). Here, we have refined this result by showing that, while the joint estimator exhibits small-sample bias and the conditional one does not, it retains its superiority in terms of mean squared error. Asymptotically, the difference between them vanishes, in both the probabilistic and deterministic cases.

Further, when $\beta = 0$, the CRSS case, the expression for both the information and for the mean squared error coincide, given that both estimators reduce to the ordinary sample average. Recall that, for this setting, the estimator is not optimal, and no uniformly optimal estimator exists, a consequence of non-completeness.

Finally, we note that the conditional likelihood estimator is also conditionally unbiased, i.e., it is unbiased for both situations $N = n$ and $N = 2n$ separately. To see this, it is convenient to rewrite

the expectation of the generalized sample average (44)

$$\begin{aligned} E(\overline{\mu}) &= c \cdot \left\{ \mu + \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{n\Phi(\nu)} \right\} \cdot \Phi(\nu) \cdot I(N=n) \\ &\quad + d \cdot \left\{ \mu - \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{2n[-1\Phi(\nu)]} \right\} \cdot [1-\Phi(\nu)] \cdot cdotI(N=2n), \end{aligned}$$

from which both expectations $E(\overline{\mu}|N)$ follow:

$$E(\overline{\mu}|N=n) = \mu + \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{n\Phi(\nu)}, \tag{56}$$

$$E(\overline{\mu}|N=2n) = \mu - \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{2n[1-\Phi(\nu)]}. \tag{57}$$

For the specific case $\beta \to \infty$, these become, either by taking limits or using expressions for truncated normals[22]:

$$E(\overline{\mu}|N=n) = \mu + \frac{\phi(\sqrt{n}\mu)}{\sqrt{n}\Phi(\sqrt{n}\mu)}, \tag{58}$$

$$E(\overline{\mu}|N=2n) = \mu - \frac{\phi(\sqrt{n}\mu)}{2\sqrt{n}[1-\Phi(\sqrt{n}\mu)]}. \tag{59}$$

Using these, it follows that $E[\widetilde{S}(\mu)|N] = 0$.

For the sample average a little more caution is required. From (56) and (57), it follows that $E(\overline{\mu}|N)$ converges to μ at a rate of n, because $\nu \to \alpha + \beta\mu$. The situation is more subtle when $\beta \to \infty$. To show this, we take the limit of (58) and (59) as $n \to \infty$. When $\mu < 0$, applying de l'Hôpital's rule whenever needed, the limits are $E(\overline{\mu}|N=n) \to 0$ and $E(\overline{\mu}|N=2n) \to \mu$. Similarly, when $\mu > 0$, the corresponding expressions are $E(\overline{\mu}|N=n) \to \mu$ and $E(\overline{\mu}|N=2n) \to \mu/2$. It follows that when $\mu = 0$, these both are equal to 0. Strictly speaking, this shows that there is bias in the conditional means. However, this needs careful qualification: when n grows large, the probability itself that $N = n$ ($N = 2n$) for negative (positive) μ goes to zero. This implies that, overall, conditional inference based on the ordinary sample average is still acceptable.

In conclusion, the sample average is conditionally and marginally biased, with the bias vanishing as n goes to infinity, except in the situations that correspond to vanishing probabilities. In contrast, the conditional estimator is unbiased, whether considered conditionally on the observed sample size or marginalized over it.

3.5 Ramifications

By starting from a simple sequential trial setting with two possible sample sizes, $N = n$ and $N = 2n$ and normally distributed outcomes, but allowing for a probabilistic stopping rule, expressed as a probit of the form $\Phi(\alpha + \beta/n \cdot k)$, we simultaneously considered three important cases: (a) when $\beta = 0$, the CRSS setting; (b) for $\beta \neq 0$ and finite, probabilistic stopping based on prior outcomes is considered, which is an example of MAR; and (c) for $\beta \to +\infty$, a conventional sequential trial.

For all three cases, the sufficient statistic (K, N) in incomplete. It follows that broad classes of estimators with the same expectation can be constructed. This is particularly noteworthy for the completely random stopping case (a), because it implies, in particular, that apart from the sample average, there is an infinite class of so-called generalized sample average estimators, without there being a uniformly optimal one. In particular, the sample average is not optimal, although it does have this property asymptotically. This is very interesting, because the sample size is an ancillary statistic in this case, so the comfort of the ancillarity property, combined with the of lack of completeness, creates at first sight counterintuitive results.

Cases (b) and (c) share a large number of properties, in the sense that there is little or no difference between the deterministic and probabilistic cases. Rather, the defining feature is the fact that "missingness" is of the MAR rather than the MCAR type. In particular, all generalized sample average estimators are biased, although a sub-class of them, including the ordinary sample average, is asymptotically unbiased.

The sample average also follows as the maximum likelihood estimator from the joint likelihood, underscoring the fact that the asymptotic bias should be zero. For the conditional likelihood, a correction term emerges that reduces the precision but removes the bias. In terms of mean squared error, the ordinary sample average is superior.

While the setting considered here is very specific in the sense that (1) there is only one intermediate look, rather than an arbitrary but fixed number, and (2) the outcomes are assumed normally distributed, the results are sufficiently generic to make the conclusions more broadly valid. Because all of the settings considered contain this probabilistic stopping rule, with the sequential trial as a special case, it follows that in all cases the simple average, combined with the (random) sample or cluster size, is an incomplete statistic. We do however need to make some further points about the more general case and we sketch this in the next section. Following that, in Sections 4 and 5 we

deal in more detail with cases (a), random sample size, and (b), missing data, respectively.

3.6 Stopping Rules with Arbitrary Numbers of Looks

We now generalize the two-sample size case, with $N = n$ or $N = 2n$, to the general $n_1 < n_2 < \ldots < n_L$, case with L a pre-specified maximum number of looks.

In line with (5), we assume that the stopping probability takes the form $F(\alpha_j + \beta_j/n_j \cdot k)$, for $j = 1, \ldots, L$. As before, if all $\beta_j = 0$, then the completely random sample size case follows, and if all $\beta_j \to +\infty$, then a sequential-trial setting follows. Based upon earlier derivations and the work by[8], it follows that

$$\begin{aligned} p_0(n_j, k) &= f_0(n_j, k) \cdot F(\alpha_j + \beta_j/n_j \cdot k), \\ f_0(n_1, k) &= \phi_{n_1}(k), \\ f_0(n_j, k) &= \int f_0(n_{j-1}, z) \cdot [1 - F(\alpha_{j-1} + \beta_{j-1}/n_{j-1} \cdot z)] \cdot \phi_{n_j - n_{j-1}}(k - z)\, dz, \end{aligned}$$

for $j = 2, \ldots, L$. A zero-expectation function $g(K, N)$ needs to satisfy

$$\sum_{j=1}^{L} e^{-\frac{1}{2} n_j \mu^2} \int g(k, n_j) \cdot p_0(n_j, k) \cdot e^{k\mu}\, dk = 0.$$

Using the same approach as in Section 3.2, such a function is obtained by choosing $g(K, N)$ for $N = n_1, \ldots, n_{L-1}$, and then computing, for $N = n_L$:

$$g(k, n_L) = -\frac{1}{p_0(k, n_L)} \sum_{j=1}^{L-1} \int g(z, n_j) \cdot p_0(n_j, z) \cdot \phi_{n_L - n_j}(k - z)\, dz.$$

One obvious choice for $N = n_1, \ldots, n_{L-1}$ is a constant $g(k, n_j) = a_j$. This establishes non-completeness in the general case.

The implications will be, as before, that the generalized sample average is biased, and this includes the ordinary sample average, except when $\beta_j = 0$ for all j, the CRSS case. The latter is important in its own right, and will be studied in the next section.

For the general sequential situation, with either a probabilistic or deterministic stopping rule, the

maximum likelihood estimators can be considered. The likelihood takes the form

$$L(\mu|k,n_\ell) = \phi_{n_\ell}(k)\cdot\prod_{j=1}^{\ell-1}\left[1-F\left(\alpha_j+\frac{\beta_j}{n_j}\cdot k\right)\right]\cdot F\left(\alpha_\ell+\frac{\beta_\ell}{n_\ell}\cdot k\right)^{I(\ell<L)}. \tag{60}$$

Let the latent variable, underlying $I(N=n_\ell)$ be T_ℓ for $\ell=1,\ldots,L-1$, defined by

$$\begin{aligned} T_\ell|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_\ell \equiv T_\ell|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_L &\sim N\left(\alpha_\ell+\frac{\beta_\ell}{n_\ell}\sum_{j=1}^{n_\ell}y_j,1\right) \\ &\sim N\left[\alpha_\ell+\frac{\beta_\ell}{n_\ell}(1\ldots1\;0\ldots0)\begin{pmatrix}y_1\\ \vdots\\ y_{n_L}\end{pmatrix},1\right]. \end{aligned}$$

This implies that, for the entire vector $\boldsymbol{T}=(T_1,\ldots,T_{L-1})'$, the conditional distribution is $\boldsymbol{T}|\boldsymbol{y}\sim N(\boldsymbol{A}+C\boldsymbol{y},I_{L-1})$ with $\boldsymbol{A}=(\alpha_1,\ldots,\alpha_{L-1})'$ and

$$C=\begin{pmatrix} 1 & \ldots & 1 & 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots \\ 1 & \ldots & 1 & 1 & \ldots & 1 & 0 & \ldots & 0 & \ldots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix},$$

an $(L-1)\times n_L$ matrix where the ℓth row starts with n_ℓ ones and ends with n_L-n_ℓ zeros. Given that $\boldsymbol{Y}\sim N\left(\mu\mathbf{1}_{n_L},I_{n_L}\right)$ it follows directly that

$$\left(\begin{array}{c}\boldsymbol{Y}\\ \hline \boldsymbol{T}\end{array}\right)\sim N\left[\left(\begin{array}{c}\mu\mathbf{1}_{n_L}\\ \hline \boldsymbol{\alpha}+\boldsymbol{\beta}\mu\end{array}\right),\left(\begin{array}{c|c} I_{n_L} & C' \\ \hline C & I_{L-1}+CC'\end{array}\right)\right].$$

Hence, the set of probabilities of observing a particular sample size take a multivariate probit form:

$$\begin{aligned} P(N=n_\ell) &= \Phi_\ell\left[\widetilde{I}_\ell\left(\boldsymbol{\alpha}^{(\ell)}+\boldsymbol{\beta}^{(\ell)}\mu\right),I_\ell+D^{(\ell)}\right], \qquad (\ell=1,\ldots,L-1) \\ P(N=n_L) &= \Phi_{L-1}\left[-\left(\boldsymbol{\alpha}+\boldsymbol{\beta}\mu\right),I_{L-1}+D\right], \end{aligned}$$

where a superscript "(ℓ)" indicates a sub-vector of the first ℓ components and $\widetilde{I}_\ell$ is a diagonal matrix with -1 along the main diagonal, except in the (ℓ,ℓ) entry, which is set to $+1$.

As a result, assuming a probit stopping rule, the conditional likelihood takes the form,

$$L_c(\mu|k,n_\ell) = \frac{\phi_{n_\ell}(k)\cdot\prod_{j=1}^{\ell-1}\left[1-\Phi\left(\alpha_j+\frac{\beta_j}{n_j}\cdot k\right)\right]\cdot\Phi\left(\alpha_\ell+\frac{\beta_\ell}{n_\ell}\cdot k\right)^{I(\ell<L)}}{\Phi_\ell\left[\widetilde{I}_\ell\left(\boldsymbol{\alpha}^{(\ell)}+\boldsymbol{\beta}^{(\ell)}\mu\right),I_\ell+D^{(\ell)}\right]^{I(\ell<L)}\cdot\Phi_{L-1}\left[-\left(\boldsymbol{\alpha}+\boldsymbol{\beta}\mu\right),I_{L-1}+D\right]^{I(\ell=L)}}.$$

The score equations and information matrices are not especially difficult to derive, but require some tedious algebra. Fitting can be done by resorting to numerical methods, possibly replacing the multivariate probit by an appropriate product of univariate probit expressions.

The arguments used earlier for the $N = n, 2n$ case to obtain the properties of the estimators based on the joint likelihood, the conditional likelihood, and generalized sample average can be extended in a straightforward way for the current more general case. We omit details, but indicate now why such a generalization follows. There are two steps. First, earlier results derived for $N = n, 2n$ carry over directly to the $N = n_1, n_2$ case. Second, when L rather than 2 sample sizes are allowed, the results can be applied sequentially to establish the key results: (a) any generalized sample average is biased, except when $\beta = 0$, although a subclass, including the ordinary sample average, will be asymptotically unbiased; (b) the joint likelihood estimator reduces to the ordinary sample average and is asymptotically unbiased; (c) the maximum conditional likelihood estimator is unbiased, yet less efficient than the joint one; (d) as we found in in Section 3.4, the joint-likelihood-based ordinary sample average has the smallest mean squared error.

4 Completely Random Sample Size

So far the CRSS case has been dealt with by setting $\beta = 0$ in the simple two sample size setup. We now extend this to the setting with sample size N with associated probability function π_n for $n = 0, \ldots, m$, and m some upper bound on the sample size, n. Evidently, $\sum_{n=0}^{m}\pi_n = 1$. When only two of these are non-zero, for n and $2n$, the earlier results are recovered.

4.1 Incomplete Sufficient Statistics

In this case, $p_0(n,k) = \phi_n(k)\pi_n$. The condition that needs to be satisfied for a function $g(k,n)$ to have zero expectation is

$$\sum_{n=0}^{m} e^{\frac{1}{2}n\mu^2}\int g(k,n)\cdot p_0(n,k)\cdot e^{\mu k}\,dk = 0.$$

The above expression can be rewritten as

$$\sum_{n=0}^{m} \pi_n \int \phi_{m-n}(k) \cdot e^{\mu k}\, dk \cdot \int g(k,n) \cdot \phi_n(k) \cdot e^{\mu k}\, dk = 0.$$

Using the uniqueness of the Laplace transform, we find

$$\sum_{n=0}^{m} \pi_n \int g(z,n) \cdot \phi_n(z) \cdot \phi_{m-n}(k-z)\, dz = 0. \tag{61}$$

Choosing $g(z,n) = b_n$, i.e., a constant not depending on k, (61) implies

$$\phi_m(k) \cdot \sum_{n=0}^{m} \pi_n \cdot b_n = 0.$$

This is possible only when

$$\sum_{n=0}^{m} \pi_n \cdot b_n = 0. \tag{62}$$

In other words, (62) can be interpreted as a vector $\boldsymbol{b}$ that is orthogonal to the probability vector $\boldsymbol{\pi}$.

4.2 Generalized Sample Average

Consider for this case a generalized sample average estimator of the form

$$\overline{\mu} = \sum_{n=0}^{m} \frac{K}{n} \cdot a_n \cdot I(N = n). \tag{63}$$

The **expectation** of (63) equals μ if and only if the vector $\boldsymbol{a}$ can be written as $\boldsymbol{a} = \mathbf{1} + \boldsymbol{b}$, with $\boldsymbol{b}$ satisfying (62) and $\mathbf{1}$ a $(m+1)$-vector of ones. In other words, the existence of an entire class of generalized sample average estimators is directly linked to the incompleteness of the sufficient statistic (K, N). One possible solution is given by $a_n = 1$, which leads to the simple sample average. This provides an additional reason to derive the variance, and explore the existence, of an optimal choice for the coefficients.

We now assume, in a slight generalization of the previous setting, that $Y_i \sim N(\mu, \sigma^2)$. The **variance** follows from the usual iterative rule and takes the form

$$\mathsf{var}(\overline{\mu}) = \mu^2 \sum_{n=0}^{m} (a_n - 1)^2 \cdot \pi_n + \sigma^2 \sum_{n=0}^{m} \frac{a_n^2}{n} \cdot \pi_n. \tag{64}$$

If $a_n = 1$, the variance reduces to

$$\mathsf{var}(\mu) = \sigma^2 \sum_{n=0}^{m} \frac{\pi_n}{n}, \tag{65}$$

as expected.

To derive the **optimal estimator**, i.e., the optimal choice for the vector $\boldsymbol{a}$, we begin by writing down the objective function

$$F(\boldsymbol{a}, \lambda) = \mu^2 \sum_{n=0}^{m} (a_n - 1)^2 \cdot \pi_n + \sigma^2 \sum_{n=0}^{m} \frac{a_n^2}{n} \cdot \pi_n - 2\lambda \left(\sum_{n=0}^{m} a_n \cdot \pi_n - 1 \right), \tag{66}$$

where λ is a Lagrange multiplier. The optimum follows from setting the derivatives of $F(\boldsymbol{a}, \lambda)$ with respect to the components a_n and λ equal to zero. First, we find that

$$a_n = \frac{1}{\mu^2 + \sigma^2/n} \cdot (\lambda + \mu^2). \tag{67}$$

Multiplying (67) by π_n and summing leads to

$$\lambda = \left(\sum_{k=0}^{m} \frac{\pi_k}{\mu^2 + \sigma^2/k} \right)^{-1} - \mu^2,$$

which in turn leads to

$$a_n = \frac{\frac{1}{\mu^2 + \sigma^2/n}}{\sum_{k=0}^{m} \frac{\pi_k}{\mu^2 + \sigma^2/k}}. \tag{68}$$

We find that the optimal generalized sample average estimator not only differs from the ordinary sample average, but is not uniform in the unknown parameters. Again, this is a consequence of incompleteness of the sufficient statistics.

In the special case that only n and $2n$ are possible sample sizes, with probabilities of occurrence Φ and $1 - \Phi$, respectively, then the earlier result (47) follows.

Throughout this section, it has been assumed that m is a finite upper limit to the sample size. It is relatively straightforward to generalize our findings to the case where N ranges over all non-negative integers. For example, N could be assumed to follow a Poisson distribution.

the stopping rule with $\beta = 0$ that was used throughout Section 3, is a special case of MCAR, whereas, if $\beta \neq 0$, the mechanism can be interpreted as MAR, in the sense that stopping, like dropout, depends on the observed outcomes, but there is no dependence on the subsequent, and potentially unobserved, outcomes. Note as well that $\beta \to \infty$ corresponds to the missing data setting, but now with a deterministic rule.

The difference from general missing-data settings, such as those arising in longitudinal studies is that, up till now, our outcomes have been univariate rather than hierarchical. This implied, as stated in Section 2, that the parameters α and β governing the stopping rule, have to be assumed known because they cannot be estimated from the data. This is the reason we now sketch a more general missing-data framework and indicate which results apply in this broader setting.

The missing data case can be expressed as follows through the generic form (3):

$$\begin{aligned} f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{\theta}^*, \boldsymbol{\psi}^*) &= f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{t}_i, \boldsymbol{\theta}^*) \cdot \frac{f(\boldsymbol{t}_i | \boldsymbol{\psi}^*)}{f(\boldsymbol{t}_i | \boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*)} \\ &= f(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m | \boldsymbol{t}_i, \boldsymbol{\theta}^*) \cdot w_i(\boldsymbol{y}_i^o, \boldsymbol{y}_i^m, \boldsymbol{t}_i, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*). \end{aligned} \tag{69}$$

Here, $\boldsymbol{t}_i$ is a vector of indicator variables the missingness status (observed: 1) of the elements of $\boldsymbol{Y}_i$. In the case of dropout in a longitudinal study, $\boldsymbol{t}_i$ can be replaced by an integer, indicating the occurrence of dropout.

When MAR holds, (69) can be replaced by an expression operating at the observed-data level only:

$$f(\boldsymbol{y}_i^o | \boldsymbol{\theta}^*, \boldsymbol{\psi}^*) = f(\boldsymbol{y}_i^o | \boldsymbol{t}_i, \boldsymbol{\theta}^*) \cdot \frac{f(\boldsymbol{t}_i | \boldsymbol{\psi}^*)}{f(\boldsymbol{t}_i | \boldsymbol{y}_i^o, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*)} = f(\boldsymbol{y}_i^o | \boldsymbol{t}_i, \boldsymbol{\theta}^*) \cdot w_i(\boldsymbol{y}_i^o, \boldsymbol{t}_i, \boldsymbol{\theta}^*, \boldsymbol{\psi}^*). \tag{70}$$

In accord with the discussion in Section 2 we need to be careful when considering the extent of ignorability. When the model is formulated in a pattern-mixture way, then the marginal observed-data distribution is governed by $\boldsymbol{\theta}^*$ and $\boldsymbol{\psi}^*$, and separability does not apply. In other words, ancillarity does not hold and the missing-data model needs to taken into account for a valid analysis. This is no longer necessary when the model is formulated in selection-model terms and under MAR, when pure likelihood or Bayesian inferences are drawn. For semi-parametric methods, inverse probability weighting can be used, noting that the missing-data model is required for this, but in this case degeneracy poses severe problems, in the sense that some weights will not be defined and the

marginal model cannot be fully recovered. For this reason the weights need to be bounded away from zero[23–24].

Example: Normally Distributed Outcomes and Probit Dropout Model

As with the sequential trial case, it is instructive to consider an example. Assume that $\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i)$ and that the dropout model takes the form:

$$f(t_i|\boldsymbol{y}_i^o) = \prod_{j=2}^{t_i-1} [1 - \Phi(\alpha_j + \beta_j y_{i,j-1})] \cdot \Phi(\alpha_{t_i} + \beta_{t_i} y_{i,d_i-1})^{I(t_i \leq n_i)}. \tag{71}$$

For simplicity, and to emphasize the connections with our initial sequential trial setting, we partition $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}', \boldsymbol{Y}_{i2}')'$, assuming sub-vector $\boldsymbol{Y}_{i1}$ is always observed and $\boldsymbol{Y}_{i2}$ is potentially missing. We partition $\boldsymbol{\mu}_i$ and Σ_i accordingly. We can now have t_i take two value only: $t_i = 1$ if the full vector is observed for subject i and $t_i = 0$ if only the first sub-vector is observed. The observed-data likelihood is

$$L = \prod_{i=1}^{N} \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_i, \Sigma_i) \cdot \Phi(\alpha + \boldsymbol{y}_{i1}\boldsymbol{\beta})^{t_i} \cdot [1 - \Phi(\alpha + \boldsymbol{y}_{i1}\boldsymbol{\beta})]^{1-t_i}. \tag{72}$$

It is convenient to introduce the latent variable Z_i that determines T_i:

$$Z_i|\boldsymbol{y}_i \sim N\left[\alpha + (\boldsymbol{y}_{i1}', \boldsymbol{y}_{i2}') \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}; 1\right].$$

Given that $\boldsymbol{Y}_i$ and Z_i are then jointly normally distributed, standard manipulations lead to

$$\left(\begin{array}{c} \boldsymbol{Y}_{i1} \\ \hline \boldsymbol{Y}_{i2} \\ \hline Z_i \end{array}\right) \sim N\left[\left(\begin{array}{c} \boldsymbol{\mu}_{i1} \\ \hline \boldsymbol{\mu}_{i2} \\ \hline \alpha + \boldsymbol{\beta}'\boldsymbol{\mu}_{i1} \end{array}\right); \left(\begin{array}{c|c|c} \Sigma_{i11} & \Sigma_{i12} & \Sigma_{i11}\boldsymbol{\beta} \\ \hline \Sigma_{i21} & \Sigma_{i22} & \Sigma_{i21}\boldsymbol{\beta} \\ \hline \boldsymbol{\beta}'\Sigma_{i11} & \boldsymbol{\beta}'\Sigma_{i12} & 1 + \boldsymbol{\beta}'\Sigma_{i11}\boldsymbol{\beta} \end{array}\right)\right], \tag{73}$$

from which the marginal probability of dropping out can be obtained

$$P(T_i = 1) = \Phi\left(\frac{\alpha + \boldsymbol{\beta}'\boldsymbol{\mu}_{i1}}{\sqrt{1 + \boldsymbol{\beta}'\Sigma_{i11}\boldsymbol{\beta}}}\right). \tag{74}$$

As in the sequential trial setting, it is interesting to consider the deterministic case, where $\boldsymbol{\beta}$ approaches infinity, in the sense that, for example, $\boldsymbol{\beta} = \lambda\boldsymbol{\beta}_0$ and $\lambda \to +\infty$. Then, the limiting

probability is

$$P(T_i = 1) = \Phi\left(\frac{\boldsymbol{\beta}_0'\boldsymbol{\mu}_{i1}}{\sqrt{\boldsymbol{\beta}_0'\Sigma_{i11}\boldsymbol{\beta}_0}}\right). \tag{75}$$

In a similar way to expression (14), the 'weight' connecting the marginal (selection model) to the pattern-mixture formulation takes the form:

$$w(\boldsymbol{y}_i^0) = \frac{\Phi\left(\frac{\alpha+\boldsymbol{\mu}_{i1}'\boldsymbol{\beta}}{\sqrt{1+\boldsymbol{\beta}'\Sigma_{i11}\boldsymbol{\beta}}}\right)}{\Phi(\alpha+\boldsymbol{y}_{i1}'\boldsymbol{\beta})}. \tag{76}$$

This 'weight' shows that, even though the selection model may be separable, the corresponding pattern-mixture model is not. Again, the degeneracy of the denominator in (76) when $\lambda \to +\infty$ underscores the parallels with the degeneracy of inverse probability weighting methods under such deterministic mechanisms.

From the above expressions, it can be seen that similar algebra establishes the properties of the maximum likelihood estimators. First, consider the joint likelihood (72) and assume the mean vector and covariance matrix are constant across units. The more general case readily follows as well. Using a partition as in (73), the maximum likelihood estimators are[1]:

$$\widehat{\boldsymbol{\mu}}_1 = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_{i1}, \tag{77}$$

$$\widehat{\boldsymbol{\mu}}_2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_{i2} - \Sigma_{21}\Sigma_{11}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_{i1} - \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_{i1}\right), \tag{78}$$

where there are N subjects all together, out of which n are fully observed and $N-n$ are only observed in the sub-vector $\boldsymbol{Y}_{i1}$.

Note that there is a fundamental difference between this and the sequential-trial case. Because of the correlation between the repeated measurements, (78) is not a simple sample average, though it could be termed a *prediction corrected sample average*. It only takes the form of a simple sample average when the first set of measurements is independent of the second, i.e., when $\Sigma_{12} = \mathbf{0}$. As in the sequential-trial case, we see that the joint estimator does not involve the missing-data model. This is again the familiar property of ignorability, applied in the selection model framework. However, the (generalized) sample average is biased, because of its failure to take $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{\theta}^*)$ into account.

Again, we consider the conditional likelihood. The counterpart to (13) is obtained through the division of (72) by the proper function of (74)

$$L_c = \frac{\prod_{i=1}^{N} \phi(\boldsymbol{y}_i^o; \boldsymbol{\mu}_i, \Sigma_i) \cdot \Phi(\alpha + \boldsymbol{y}_{i1}\boldsymbol{\beta})^{t_i} \cdot [1 - \Phi(\alpha + \boldsymbol{y}_{i1}\boldsymbol{\beta})]^{1-t_i}}{\Phi\left(\frac{\alpha+\boldsymbol{\beta}'\boldsymbol{\mu}_{i1}}{\sqrt{1+\boldsymbol{\beta}'\Sigma_{i11}\boldsymbol{\beta}}}\right)^{t_i} \cdot \left[1 - \Phi\left(\frac{\alpha+\boldsymbol{\beta}'\boldsymbol{\mu}_{i1}}{\sqrt{1+\boldsymbol{\beta}'\Sigma_{i11}\boldsymbol{\beta}}}\right)\right]^{1-t_i}} \tag{79}$$

Although we may be primarily interested in the parameters governing $\boldsymbol{\mu}$ and/or the covariance parameters, the missing-data mechanism will enter the likelihood equations through contributions based on the denominator of (79). This likelihood is of the pattern-mixture form: as seen earlier, ignorability is model-framework specific and the fact that we are be estimating selection-model parameters through a pattern-mixture likelihood makes the missing data mechanism non-ignorable, even under MAR. It can be shown for this setting as well (details not given) that, while the conditional estimator is unbiased, whereas the marginal one is only asymptotically unbiased, the latter has the smaller mean squared error.

6 Concluding Remarks

The developments in this paper show that, when the sample size is random, an estimator as familiar as the sample average has frequentist properties that are not altogether straightforward. This arises from the interplay between the properties of (non-)ancillarity and incompleteness.

In the case of a completely random sample size, the sample average is unbiased but not optimal. It is one member of a class of generalized sample averages that is unbiased. This class no uniformly optimal member. However, the sample average does converge to the optimum for increasing sample size. The ordinary average is unique in having weights that do not change regardless of when exactly the sequence is stopped; this is definitely an advantage. In the light of these results, the sample average remains the most sensible practical choice.

In the cases of univariate incomplete data and sequential trials, the sample average is biased in small samples but asymptotically unbiased, both conditionally and marginally. This follows from direct frequentist calculations, as well as from invoking likelihood theory. The exact calculations done here allow us to gauge the extent of the bias. The sample average follows as the maximum likelihood estimator from the joint likelihood for outcomes and sample size. In spite of confusion about the sample size estimator in the literature, this result is not totally surprising, given that it

follows from invoking ignorability in the likelihood context. By contrast, the maximum conditional likelihood estimator is unbiased but the ordinary sample average still has the smaller mean squared error. This estimator can be interpreted as the one based on the pattern-mixture decomposition of the likelihood. These results combined together indicate that the use of the ordinary sample average in a sequential trial is a sensible choice, provided an estimator of precision based on the observed information be used[8–12–19].

The majority of our derivations has been in terms of specific assumptions, namely a probabilistic stopping rule of a probit form, standard normally distributed outcomes, and two possible sample sizes, n and $2n$. Extensions to (1) arbitrary sample sizes in the completely random sample size settings, (2) general sequential trials with $L > 1$ looks, (3) normal outcomes with other than unit variance, and (4) more general missing data problems have also been considered. In (4) it has been shown that the ordinary sample average is no longer (asymptotically) unbiased because the joint likelihood estimator, again following from invoking ignorability under MAR, involves the predictive mean of the unobserved outcomes, given the observed ones. This well known correction[1] is a direct consequence of the correlation between observed and unobserved measurements.

Although the probit form for the stopping rule/missing data mechanism and the normality of the outcomes may be seen as limitations, the developments are generic. The advantage of our choices is the existence of explicit expressions that permit insightful interpretation. These choices are also practically relevant for many settings.

The results developed here have important practical implications for a wide range of settings. For sequential trials, there has been long-standing confusion and controversy regarding the (in)appropriateness of the sample average when estimating a parameter after such a trial. Our results show that the ordinary sample average, while small-sample biased and not uniform minimum variance unbiased, is perfectly acceptable. This should be seen against the background of the conditional likelihood estimator. Even though the latter is small sample unbiased, it suffers from a slightly increased mean square error. Thus, in conclusion, while some familiar properties no longer hold, estimation after sequential trials is more straightforward than commonly considered and there is less need for complicated, modified estimators than perhaps generally thought, given that the ordinary sample average can be used without trouble. While our results were, for clarity, presented for a fairly simple setting, they carry over for general studies, with multiple looks, different types of stopping rules, etc.

In the same vein, many clinical studies are prone to incompleteness. Our results provide a bridge between them. Hence, very similar considerations apply. The main difference between both is that incomplete data frequently occur in a setting where several measurements are collected on the same subject and that the missing-data mechanism is generally, but not always, stochastic rather than deterministic. Our derivations show that these differences are not fundamental and nicely connect the likelihood-based ignorability theory to the validity of the simple sample average in the sequential trial setting.

Further work will draw out similar results for studies with random cluster sizes, and for studies where longitudinal and time-to-event data are collected simultaneously. Also the missing-data case will be scrutinized further.

Our results have implications for several broad settings not considered in detail here. Examples are clusters with random size, censored time-to-event outcomes, joint modeling of longitudinal and survival outcomes, and random observation times. These will be the subject of further research.

Acknowledgments

Geert Molenberghs, Mike Kenward, Marc Aerts, and Geert Verbeke gratefully acknowledge support from IAP research Network P6/03 of the Belgian Government (Belgian Science Policy). The work of Anastasios Tsiatis and Marie Davidian was supported in part by NIH grants P01 CA142538, R37 AI031789, R01 CA051962, and R01 CA085848.

Table 1: *Simulation study for the sequential trial case. n: sample size; μ: true mean for the standard normal from which the sample is taken; α, β: parameters governing the stopping rule; #stop: number of simulations out of 1000 for which stopping is applied; $\widehat{\mu}$: average mean for the joint approach; $\widehat{s}$: average standard error for the joint approach; $\widehat{\mu}_c$: average mean for the conditional approach; $\widehat{s}_c$: average standard error for the conditional approach.*

n	μ	α	β	#stop	$\widehat{\mu}$	$\widehat{s}$	$\widehat{\mu}_c$	$\widehat{s}_c$
10	0	0	0	486	0.018488	0.268621	0.018488	0.268621
100	0	0	0	516	-0.004514	0.085824	-0.004514	0.085824
1000	0	0	0	484	-0.000460	0.026844	-0.000460	0.026844
10	0	0	1	488	0.019749	0.268806	0.001946	0.275121
100	0	0	1	477	0.002681	0.084682	0.000972	0.084891
1000	0	0	1	488	0.000375	0.026881	0.000190	0.026887
10	0	1	0	849	0.007495	0.302242	0.007495	0.302242
100	0	1	0	846	-0.000064	0.095489	-0.000064	0.095489
1000	0	1	0	840	0.001791	0.030141	0.001791	0.030141
10	1	0	0	525	0.989761	0.272233	0.989761	0.272233
100	1	0	0	519	0.998925	0.085912	0.998925	0.085912
1000	1	0	0	515	0.999771	0.027131	0.999771	0.027131
10	1	0	1	809	1.014183	0.298537	1.003231	0.303925
100	1	0	1	844	0.996496	0.095431	0.995216	0.095609
1000	1	0	1	854	1.000929	0.030271	1.000794	0.030276
10	1	1	0	832	1.014265	0.300667	1.014265	0.300667
100	1	1	0	845	0.996084	0.095460	0.996084	0.095460
1000	1	1	0	842	0.998763	0.030159	0.998763	0.030159
10	-1	0	0	515	-1.006709	0.271307	-1.006709	0.271307
100	-1	0	0	493	-1.001493	0.085150	-1.001493	0.085150
1000	-1	0	0	501	-1.000682	0.027001	-1.000682	0.027001
10	-1	0	1	171	-1.001404	0.239445	-1.014000	0.243141
100	-1	0	1	151	-0.995329	0.075133	-0.996375	0.075249
1000	-1	0	1	140	-0.999724	0.023657	-0.999813	0.023661
10	-1	1	0	840	-1.005741	0.301408	-1.005741	0.301408
100	-1	1	0	866	-0.997533	0.096075	-0.997533	0.096075
1000	-1	1	0	842	-0.998102	0.030159	-0.998102	0.030159
10	0	0	10	479	0.050863	0.267972	-0.078807	0.400683
100	0	0	10	511	0.014664	0.085678	-0.004165	0.099474
1000	0	0	10	499	0.001780	0.026982	-0.000150	0.027620
10	1	0	10	1000	1.003133	0.316228	0.988938	0.328269
100	1	0	10	1000	1.000227	0.100000	1.000227	0.100000
1000	1	0	10	1000	1.000494	0.031623	1.000494	0.031623
10	-1	0	10	1	-1.004547	0.223699	-1.002906	0.225618
100	-1	0	10	0	-0.998418	0.070711	-0.998418	0.070711
1000	-1	0	10	0	-0.999392	0.022361	-0.999392	0.022361

On Random Sample Size, Ignorability, Ancillarity, Completeness, Separability, and Degeneracy: Sequential Trials, Random Sample Sizes, and Missing Data

Geert Molenberghs[1,2] **Michael G. Kenward**[3] **Marc Aerts**[1] **Geert Verbeke**[2,1]

Anastasios A. Tsiatis[4] **Marie Davidian**[4] **Dimitris Rizopoulos**[5]

[1] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[2] *I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

[3] *Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London WC1E7HT, United Kingdom*

[4] *Department of Statistics, North Carolina State University, Raleigh, NC, U.S.A.*

[5] *Department of Biostatistics, Erasmus University Medical Center, NL-3000 CA Rotterdam, the Netherlands*

Appendix

A Completeness, Unbiasedness, and Minimum Variance Calculations for the Stopping Rule Cases

Details behind key derivations in Section 3 are given here.

A.1 Derivation of $p_0(2n,k)$

The function takes the form:

$$p_0(2n,k) = \phi_{2n}(k) - \int \phi_n(z)\cdot\phi_n(k-z)\cdot\Phi(\alpha+\beta z/n)\,dz = \phi_{2n}(k) - \widetilde{p}_0(2n,k).$$

Write

$$p_0(2n,k) = \frac{1}{(2\pi n)^{3/2}}\int_{z=-\infty}^{z=+\infty}\int_{t=-\infty}^{t=\alpha+\beta/n\cdot z}\exp\left[-\frac{1}{2}\frac{z^2}{n} - \frac{1}{2}\frac{(k-z)^2}{n} - \frac{1}{2}t^2\right]\,dt\,dz$$

and apply the change of variables: $t = \beta/nz + s$. The integral becomes

$$\widetilde{p}_0(2n,k) = \frac{1}{(2\pi)^{3/2}n}\int_{s=-\infty}^{s=\alpha} e^{-\frac{1}{2n}p}\,ds\cdot\int_{z=-\infty}^{z=+\infty}\exp\left[-\frac{1}{2n}\left(\sqrt{\frac{2n+\beta^2}{n}}z+m\right)^2\right]\,dz$$

with

$$\begin{aligned} m &= \sqrt{\frac{2n+\beta^2}{n}}(s\beta - k), \\ p &= \frac{2n^2s^2 + 2n\beta ks + k^2(n+\beta^2)}{2n+\beta^2}. \end{aligned}$$

Upon rearranging terms, we obtain:

$$\widetilde{p}_0(2n,k) = \frac{\sqrt{n}}{2\pi\sqrt{n}\sqrt{2n+\beta^2}} e^{-\frac{1}{2n}\frac{k^2(n+1/2\beta^2}{2n+\beta^2}} \int_{s=-\infty}^{s=\alpha} \exp\left[-\frac{1}{2n}\frac{(\sqrt{2}ns + \sqrt{2}/2\beta k)^2}{2n+\beta^2}\right] ds.$$

Applying a further change of variables

$$q = \frac{s + \frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}}$$

the integral reduces to

$$\widetilde{p}_0(2n,k) = \phi_{2n}(k) \cdot \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}}\right).$$

A.2 Expectation of the Generalized Sample Average

To show that the expectation of (41) is equal to (44), it is useful to first consider the auxiliary quantity:

$$\begin{aligned} I &= \int k \cdot f_N(k) \cdot \Phi\left(A + \frac{B}{N}k\right) dk \qquad (A.1) \\ &= \frac{1}{2\pi\sqrt{N}} \sum_{k=-\infty}^{k=+\infty} dk \int_{t=-\infty}^{t=A+\frac{B}{N}k} dt\, k \cdot \exp\left[-\frac{1}{2N}(k-N\mu)^2 - \frac{1}{2}t^2\right] \\ &= \frac{1}{\sqrt{2\pi}} \int_{s=-\infty}^{s=A} \exp\left[-\frac{1}{2+2B^2/N}(s+\mu B)^2\right] ds, \end{aligned}$$

upon applying the change of variables and $t = s + B/N \cdot k$ and completing the square, we obtain:

$$\begin{aligned} I &= \frac{1}{\sqrt{2\pi}} \frac{1}{(1+B^2/N)^{3/2}} \left\{ N\mu \int_{s=-\infty}^{s=A} \exp\left[-\frac{1}{2}\left(\frac{s+\mu B}{\sqrt{1+B^2/N}}\right)^2\right] ds \right. \\ &\qquad \left. -B \int_{s=-\infty}^{s=A} s \cdot \exp\left[-\frac{1}{2}\left(\frac{s+\mu B}{\sqrt{1+B^2/N}}\right)^2\right] ds \right\}. \\ &= N\mu\Phi\left(\frac{A+B\mu}{\sqrt{1+B^2/N}}\right) + \frac{B}{\sqrt{2\pi}\sqrt{1+B^2/N}} \exp\left[-\frac{1}{2}\left(\frac{A+B\mu}{\sqrt{1+B^2/N}}\right)\right]. \end{aligned}$$

Returning to (41), it follows that

$$E(\overline{\mu}) = \frac{c}{n}\int k\cdot f_n(k)\cdot\Phi\left(\alpha+\frac{\beta}{n}k\right)\,dk + \frac{d}{2n}\int k\cdot f_{2n}(k)\cdot\left[1-\Phi\left(\frac{\alpha+\frac{\beta}{2n}k}{\sqrt{\frac{2n+\beta^2}{2n}}}\right)\right]\,dk$$
$$= d\cdot\mu+\frac{1}{2n}(2cI_1+dI_2). \tag{A.2}$$

Now, both I_1 and I_2 are of the form (A.1) with, for I_1: $N=n$, $A=\alpha$, and $B=\beta$ and, for I_2: $N=2n$, $A=\alpha/\sqrt{\frac{2n+\beta^2}{2n}}$, and $B=\beta/\sqrt{\frac{2n+\beta^2}{2n}}$. As a consequence, (A.2) can be written as:

$$E(\overline{\mu}) = d\cdot\mu+(c-d)\Phi(\nu)+\frac{2c-d}{2n\sqrt{2\pi}}\frac{\beta}{\sqrt{1+\beta^2/n}}e^{-\frac{1}{2}\nu^2},$$

with $\nu=(\alpha+\beta\mu)/\sqrt{1+\beta^2/n}$. This result is based upon the fact that, for both I_1 and I_2, the identities hold: $A/\sqrt{1+B^2/N}=\alpha/\sqrt{1+\beta^2/n}$ and $B/\sqrt{1+B^2/N}=\beta/\sqrt{1+\beta^2/n}$. Therefore, (44) follows.

A.3 Variance for Generalized Sample Average When $\beta=0$

To derive (46), observe:

$$E(\overline{\mu}|N=n) = c\mu,$$
$$E(\overline{\mu}|N=2n) = \frac{1-c\Phi}{1-\Phi},$$

from which it follows that:

$$\text{var}[E(\overline{\mu}|N)] = \mu^2(1-c)^2\frac{\Phi}{1-\Phi}. \tag{A.3}$$

Further,

$$\text{var}(\overline{\mu}|N=n) = \frac{c}{n},$$
$$\text{var}(\overline{\mu}|N=2n) = \frac{(1-c\Phi)^2}{2n(1-\Phi)^2},$$

producing

$$E[\text{var}(\overline{\mu}|N)] = \frac{\Phi\cdot c^2}{n}+\frac{(1-c\Phi)^2}{2n(1-\Phi)}. \tag{A.4}$$

Adding (A.3) and (A.4) produces (46).

Calculating the derivative of (46) with respect to c and equating it to zero yields (47). It is easy to verify that this optimum corresponds to a minimum.

References

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002.

2. Wald A. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics* 1945; **16**: 117-86.

3. Grambsch P. Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Annals of Statistics* 1983; **11**: 68–77.

4. Barndorff-Nielsen O, Cox DR. The effect of sampling rules on likelihood statistics. *International Statistical Review* 1984; **52**: 309–26.

5. Siegmund D. Estimation following sequential tests. *Biometrika* 1978; **64**: 191–99.

6. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* 1988; **7**: 1231–42.

7. Emerson SS, Fleming TR. Parameter estimation following group sequential hypothesis testing. *Biometrika* 1990; **77**: 875–92.

8. Liu A, Hall WJ. Unbiased estimation following a group sequential test. *Biometrika* 1999; **86**: 71–8.

9. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581–92.

10. Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; **88**: 125–34.

11. Little RJA. A class of pattern-mixture models for normal incomplete data. *Biometrika* 1994; **81**: 471–83.

12. Kenward MG, Molenberghs G. Likelihood based frequentist inference when data are missing at random. *Statistical Science* 1998; **13**: 236–247.

13. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman & Hall, 1974.

14. Casella G, Berger RL. *Statistical Inference*. Duxbury Press, 2001.

15. Basu D. On statistics independent of a complete sufficient statistic. *Sankhya* 1955; **15**: 377–80.

16. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**: 663–85.

17. Armitage P. *Sequential Medical Trials*. Blackwell, 1975.

18. Tsiatis AA, Rosner GL, Mehta CR Exact confidence intervals following a group sequential test. *Biometrics* 1984; **40**: 797–803.

19. Rosner GL, Tsiatis AA. Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* 1988; **75**: 723–29.

20. Todd S, Whitehead J, Facey KM. Point and interval estimation following a sequential clinical trial. *Biometrika* 1996; **83**: 453–61.

21. Whitehead J. A unified theory for sequential clinical trials. *Statistics in Medicine* 1999; **18**: 2271–86.

22. Patel JK, Read CB. *Handbook of the Normal Distribution*. Marcel Dekker, 1996.

23. Rotnitzky A. Inverse probability weighted methods. In: *Longitudinal Data Analysis* (G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs eds.), 453–476. CRC/Chapman & Hall, 2009.

24. Vansteelandt S, Carpenter JR, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology* 2010: **6**: 37–48.