

Properties of Estimators in Exponential Family Settings with
Observation-based Stopping Rules

Peer-reviewed author version

MILANZI, Elasma; MOLENBERGHS, Geert; ALONSO ABAD, Ariel; VERBEKE, Geert; Kenward, Michael G.; Tsiatis, Anastasios A. & Davidian, Marie (2016) Properties of Estimators in Exponential Family Settings with Observation-based Stopping Rules. In: Journal of Biometrics & Biostatistics, 7(1), ART N° 272.

DOI: 10.4172/2155-6180.1000272

Handle: <http://hdl.handle.net/1942/23432>

Properties of Estimators in Exponential Family Settings With Observation-based Stopping Rules

Elasma Milanzi¹ Geert Molenberghs^{1,2} Ariel Alonso³

Michael G. Kenward⁴ Geert Verbeke^{2,1}

Anastasios A. Tsiatis⁵

Marie Davidian⁵

¹ *I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium*

² *I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium*

⁵ *Department of Methodology and Statistics, Maastricht University, the Netherlands*

⁴ *Department of Medical Statistics, London School of Hygiene and Tropical Medicine,
London WC1E7HT, United Kingdom*

⁵ *Department of Statistics, North Carolina State University, Raleigh, NC, U.S.A.*

Abstract

Often, sample size is not fixed by design. A typical example is a sequential trial with a stopping rule, where stopping is based on what has been observed at an interim look. While such designs are used for time and cost efficiency, and hypothesis testing theory has been well developed, estimation following a sequential trial is a challenging, controversial problem. Progress has been made in the literature, predominantly for normal outcomes and/or for a deterministic stopping rule. Here, we place these settings in the broader context of outcomes following an exponential family distribution and, with a stochastic stopping rule that includes a deterministic rule and completely random sample size as special cases. We study (1) the so-called incompleteness property of the sufficient statistics, (2) a general class of linear estimators, and (3) joint and conditional likelihood estimation. Apart from the general exponential family setting, normal and binary outcomes are considered as key examples. While our results hold for a general number of looks, for ease of

exposition we focus on the simple yet generic setting of two possible sample sizes, $N = n$ or $N = 2n$.

Some Keywords: Completely Random Sample Size; Frequentist Inference; Generalized Sample Average; Stochastic Stopping Rule; Joint Modeling; Likelihood Inference; Missing at Random.

1 Introduction

It is commonly known that statistical designs where the sample size is random pose challenges beyond the fixed sample-size case and that many findings are counter-intuitive. While this has been documented for situations where the sample size depends on the data, such as in sequential trials (Siegmund, 1978; Hughes and Pocock, 1988; Emerson and Fleming, 1990) or incomplete data (Little and Rubin, 2002), it is less widespread that such counterintuitive results apply even when the sample size is completely random (Grambsch, 1983; Barndorff-Nielsen and Cox, 1984), in the sense that both the collected and uncollected data have no relationship to the stochastic mechanism governing the sample size. Liu and Hall (1999) provided a general theory for sequential studies, where the decision to either stop or continue the study at every interim look depends deterministically on the data collected up to that point. Molenberghs *et al* (2012) generalized their results to the setting where the sample size may depend stochastically rather than deterministically on the observed data, a general setting that contains both sequential trials and completely random sample sizes (CRSS) as special cases. We refer to these three settings together as a stochastic stopping rule. Molenberghs *et al* (2012) also discussed the related cases of incomplete longitudinal data, censored time-to-event data, joint modeling of survival and longitudinal data, and clustered data with random cluster sizes.

An important finding of Liu and Hall (1999) was that the commonly used sufficient statistics in deterministic stopping designs are incomplete, a property that will be defined in the next section. Molenberghs *et al* (2012) generalized this to stochastic stopping

rules and explore the implications of this for linear estimators based on the sample sum as well as on so-called marginal and conditional estimators. They found for stochastic stopping rules that the counterintuitive implications of a random sample size follows from two properties: (a) excluding the CRSS case, the sample size is *non-ancillary* given the sample sum; (b) the pair consisting of the accumulating sample sum and the sample size is an *incomplete* minimal sufficient statistic. These properties are defined in Section 2.

The work of Liu and Hall (1999) and Molenberghs *et al* (2012) was confined to the special case of normally distributed outcomes. Further, Molenberghs *et al* (2012) illustrated there developments with a random stopping rule of probit form. These specific choices allow for insightful expressions. The latter choice is not however necessary for deterministic stopping rules that can be cast in the form of continuation and stopping regions or, equivalently, the boundaries between them.

Extending the results in Liu and Hall (1999), Liu *et al* (2006) presented a general deterministic stopping rule theory where the outcome follows a one-parameter exponential family, and also established incompleteness for this case. This implies, in particular, that there are infinitely many unbiased estimators, none with uniformly minimum variance. Here, we show incompleteness in the exponential family case, for a stochastic stopping rule, and derive explicit results for linear estimators as well as for marginal and conditional likelihood estimators. These general findings are then further illustrated in the normal case, making the connection to Molenberghs *et al* (2012), and in the case where the outcomes are binary, and hence the sample sum is binomial.

Our findings are essentially as follows. The classical sample average is biased in finite samples, though asymptotically unbiased for a broad classes of stopping rules. An unbiased estimator follows from the conditional likelihood, where the conditioning is on the (non-ancillary) sample size. Contrary to intuition, the conditional estimator has larger mean squared error than the ordinary sample average for sufficiently large sample size,

the latter resulting from the joint likelihood, where ‘joint’ means a simultaneous model for the outcomes and the sample size. In some cases, the result holds for all sample sizes, large and small. Thus, the sample average is a valid and sensible estimator, contrary to some claims in the sequential-trial literature, for stochastic and deterministic stopping rules. The literature on sequential trials is indeed very large, with a relatively early review given by Whitehead (1999). Tsiatis, Rosner, and Mehta (1984) and Rosner and Tsiatis (1988) address precision estimation after group sequential trials. Emerson and Fleming (1990) propose estimators within an ordering paradigm. Much of this work is placed in a unifying framework by Liu and Hall (1999). A review can be found in Molenberghs *et al* (2012).

The finite-sample bias in the sample average disappears 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 for most exponential-family members; the exponential distribution is a noteworthy exception.

The case of two possible sample sizes, $N = n$ and $N = 2n$ is simple yet generic, and will be adopted here. All developments can be generalized with ease to the setting with L possible sample sizes and accrual numbers n_1, \dots, n_L .

The remainder of this paper is organized as follows. In Section 2, the problem under investigation is formally introduced, along with key concepts. The incompleteness of the sufficient statistics is established in Section 3. Section 4 is devoted to generalized sample averages, while joint and conditional likelihood estimation is the topic of Section 5. In each of Sections 3–5, the general exponential family case is supplemented with the particular case of the normal and Bernoulli distributions.

2 Notation, Basic Concepts, and Problem Formulation

As stated in the introduction, we consider a simple sequential trial, where n measurements Y_i are observed, after which a stochastic stopping rule is applied and, depending on the outcome, another set of n measurements is or is not observed. Let \mathbf{Y} be the $(2n \times 1)$ vector of outcomes that could be collected, with the sample sum denoted by K , and N be the realized sample size, that is, $N = n$ or $N = 2n$. A joint model for the stochastic outcomes is

$$f(\mathbf{y}, N|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(N|\mathbf{y}, \boldsymbol{\psi}) \quad (1)$$

$$= f(\mathbf{y}|N, \boldsymbol{\theta}, \boldsymbol{\psi}) \cdot f(N|\boldsymbol{\theta}, \boldsymbol{\psi}). \quad (2)$$

The sample sum is denoted by K . If necessary, a subscript will indicate over which batch the sample is calculated. Molenberghs *et al* (2012) noted the similarity with missing-data concepts, where (1) is a selection model factorization and (2) is a pattern-mixture factorization (Little and Rubin, 2002). In all cases, it is assumed that $f(N|\mathbf{y}, \boldsymbol{\psi}) = f(N|\mathbf{y}^o, \boldsymbol{\psi})$ depends on observed outcomes only, and hence the sample size is determined by the first batch of observations (Y_1, \dots, Y_n) . We may then write $f(N|K_n, \boldsymbol{\psi})$. This corresponds to the frequentist concept of missingness at random (Little and Rubin, 2002). In the limiting case of a deterministic stopping rule, $f(N|\mathbf{y}, \boldsymbol{\psi})$ is degenerate and $f(N = n|\mathbf{y}, \boldsymbol{\psi})$ equals 1 when $K_n \in \mathcal{S} \subset \mathbb{R}$ and 0 over its complement \mathcal{C} , with the reverse holding for $f(N = 2n|\mathbf{y}, \boldsymbol{\psi})$. The CRSS case follows by assuming \mathbf{Y} and N to be independent, meaning that both factorizations (1) and (2) trivially reduce to $f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(N|\boldsymbol{\psi})$.

In the stopping-rule case $\boldsymbol{\psi}$ is not estimable from the data and will be assumed to be specified by design. This is different for the other settings that can also be cast in terms of (1)–(2), such as incomplete longitudinal data, clusters of random size, censored time-to-event data, joint models for longitudinal and time-to-event data, and random

measurement times settings, as noted by Molenberghs *et al* (2012). In these cases, a subject-specific index i needs to be introduced into (1)–(2) and N needs to be replaced by the missing data indicators, censoring indicators, and so on.

2.1 Basic Concepts

In line with Molenberghs *et al* (2012), we will review several fundamental concepts that are essential in what follows.

In line with Rubin (1976), we consider *ignorability*. For pure likelihood or Bayesian inferences, under missingness at random (MAR), inferences about $\boldsymbol{\theta}$ can be made using $f(\mathbf{y}_i^o|\boldsymbol{\theta})$ only, without the need for an explicit missing-data mechanism or, in our case, without the need for an explicit sample-size model. This is, provided the regularity condition of *separability* holds true, i.e., that the parameter space of $(\boldsymbol{\theta}', \boldsymbol{\psi}')$ is the Cartesian product of their individual product spaces. In other words, this means that the sample size model does not contain information about the outcome model parameter. It implies that N could then be considered ancillary in the sense of (Cox and Hinkley, 1974, pp. 32–35). We will see that this is true for CRSS, but not for the other situations. Excluding MNAR, ignorability can be violated in three ways. First, even in the likelihood and Bayesian frameworks and under MAR, ignorability does not apply in a non-separable situation. Second, frequentist inferences are not necessarily ignorable under MAR. Third, assuming MAR and separability hold and we are in a likelihood or Bayesian framework, 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). The latter statement is symmetric and could be made starting from a pattern-mixture view as well. The bottom line is that ignorability holds in at most one of these, except in the trivial MCAR setting, such as for CRSS.

There is a connection between ignorability and *ancillarity* (Cox and Hinkley, 1974). They

define an ancillary statistic T to be one that complements a minimally sufficient statistic S such that, given S , T does not contain information about the parameter of interest. Arguably the best known example is the sample size $T = n$ when estimating a mean, provided the sample size is fixed by design or the law governing it does not depend on the mean parameter to be estimated, as with CRSS. Counterexamples are the stochastic and deterministic stopping rules.

The crucial property for Liu and Hall (1999), Liu *et al* (2006), Molenberghs *et al* (2012), as well as for us here is that of *completeness* (Casella and Berger, 2001, 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 for us surfaces in two ways. First, from the Lehman-Scheffé theorem (Casella and Berger, 2001), if a statistic is unbiased, complete, and sufficient for some parameter θ , then it is the best mean-unbiased estimator for θ . The lack of this property in the stopping-rule case will manifest itself when studying generalized sample averages in Section 4. Second, completeness and ancillarity are connected through Basu’s theorem (Basu, 1955; Casella and Berger, 2001, p. 287): a statistic both complete and sufficient is independent of any ancillary statistic.

2.2 General Model Formulation

Assume that we collect n i.i.d. observations Y_1, \dots, Y_n , with exponential family density

$$f_\theta(y) = h(y) \exp \{ \theta y - a(\theta) \}, \quad (3)$$

where θ is the natural parameter, $a(\theta)$ the mean generating function, and $h(y)$ a normalizing constant. Assume a stochastic stopping rule

$$\pi(N = n | k_n) = F(k_n | \boldsymbol{\psi}) = F(k_n), \quad (4)$$

with $K_n = \sum_{i=1}^n Y_i$. The form for (4) is left unspecified at this time. The CRSS setting follows as $F(k_n) \equiv F$, a constant. Likewise, when $F(\cdot)$ is degenerate, a deterministic

stopping rule ensues. When the trial is not stopped, a further n observations Y_{n+1}, \dots, Y_{2n} are collected, also with density (3). The inferential goal is to estimate θ or a function of this, such as the population mean μ . From the exponential-family structure, the density of K_n can be expressed

$$f_{n,\theta}(k) = h_n(k) \exp \{ \theta k - na(\theta) \}. \quad (5)$$

When no ambiguity can arise, the subscript n may be dropped from K_n . Because the density integrates to 1, it trivially follows that

$$e^{na(\theta)} = \int h_n(k) e^{\theta k} dk = \mathcal{L} \{ h_n(k) \}. \quad (6)$$

While expression (6) is well known to be a Laplace transformation, it is useful to state it explicitly in preparation of the derivations in Section 3. Because the stopping rule depends on K_n , and because (4) combined with the outcome model is a pattern-mixture factorization (2), N is not ancillary to K .

When, in addition, the conditional probability of stopping an exponential family form is chosen, e.g.,

$$F(k_n) = F(k) = \int_{z=-\infty}^{z=A(k)} \tilde{f}_1(z) dz, \quad (7)$$

then an appealing form for the marginal stopping probability can be derived. Here $\tilde{f}_1(z)$ can be seen as an exponential family member, underlying the stopping process. When the outcomes Y and hence K do not range over the entire real line, the lower integration limit in (7) should be adjusted accordingly, and the function $A(k)$ should be chosen so as to obey the range restrictions. It is convenient to assume that $\tilde{f}_1(z)$ has no free parameters; should there be the need for such, then they can be absorbed into $A(k)$. Hence, we can write

$$\tilde{f}_1(z) = \tilde{h}_1(z) \exp \{ -\tilde{a}(0) \}. \quad (8)$$

Using (5) and (8), the marginal stopping probability becomes:

$$\begin{aligned}
P(N = n) &= \int_{k=-\infty}^{k=+\infty} \int_{z=-\infty}^{z=A(k)} f_{n,\theta}(k) \tilde{f}_1(z) dz dk \\
&= \exp \{-na(\theta) - \tilde{a}(0)\} \int_{k=-\infty}^{k=+\infty} h_n(k) \left[\int_{z=-\infty}^{z=A(k)} \tilde{h}_1(z) dz \right] e^{\theta k} dk \\
&= \exp \{-na(\theta) - \tilde{a}(0)\} \mathcal{L} \{H_1(A(k)) \cdot h_n(k)\}, \tag{9}
\end{aligned}$$

where

$$H_1(t) = \int_{z=-\infty}^{z=t} \tilde{h}_1(z) dz.$$

In the special case of a CRSS, $A(k) \equiv A$ and (9) reduces to

$$\begin{aligned}
f(N = n) &= \exp \{-na(\theta) - \tilde{a}(0)\} H_1(A) \mathcal{L} \{h_n(k)\} \\
&= \exp \{-na(\theta) - \tilde{a}(0)\} H_1(A) e^{na(\theta)} = \int_{k=-\infty}^A \tilde{f}_1(k) dk.
\end{aligned}$$

In our two special cases, (3) will be chosen as standard normal and Bernoulli, respectively.

In the first of these, in concordance with Molenberghs *et al* (2012), (4) will be assumed to be of probit form:

$$F(k) = \Phi \left(\alpha + \beta \frac{k}{n} \right). \tag{10}$$

In the binary case, we will generally leave (4) unspecified, but for some developments it is useful to consider an explicit example, for which we will resort to the beta distribution, i.e.,

$$\tilde{f}_1(z) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha, \beta)}, \tag{11}$$

with $B(\cdot, \cdot)$ the beta function. It is convenient to choose integer values, for illustrative purposes: $\alpha = p + 1$, $\beta = q + 1$, with p and q integers, changing (11) to:

$$\tilde{f}_1(z) = (p + q + 1) \binom{p + q}{p} z^p (1 - z)^q. \tag{12}$$

Choosing (12) leads to the conditional stopping probability:

$$F(k) = (p + q + 1) \binom{p + q}{p} \sum_{\ell=0}^q \frac{(-1)^\ell}{p + \ell + 1} \binom{q}{\ell} A(k)^{p+\ell+1}. \tag{13}$$

It is instructive to consider some special cases of this. When $p = q = 0$, (12) reduces to the uniform distribution on the unit interval, and it immediately follows that $F(k) = A(k)$. When $p = 1$ and $q = 0$, we find $F(k) = A(k)^2$. As a third and last instance, when $p = q = 1$, $F(k) = 3A(k)^2 - 2A(k)^3$.

A useful function is $A(k) = k/n$, implying that stopping is certain when $K = n$ and continuation is certain when $K = 0$, while for $0 < K < n$ stopping is probabilistic. The actual probability in these cases depends on the choice for p and q .

These choices are made to illustrate our general developments and our emphasis is not on, say, designing a particular trial. However, the class of beta-based stopping rules, for example, potentially leads to rich families of stopping rules and spending functions (Whitehead, 1997; Jennison and Turnbull, 2000).

3 Incomplete Sufficient Statistics

3.1 The General Case

We now consider the role of completeness in this setting, building upon the work of Liu and Hall (1999), Liu *et al* (2006), and Molenberghs *et al* (2012). A sufficient statistic for this setting is (K, N) . In line with the developments in the above papers, the joint distribution for (K, N) is:

$$p(k, n) = f_n(k) \cdot F(k), \tag{14}$$

$$p(k, 2n) = f_{2n}(k) - \int f_n(z) f_n(k - z) F(z) dz. \tag{15}$$

When the stopping rule leads to range restrictions in the sense of Lehman and Stein (1950), it is known that the sufficient statistic is complete. Hence, for the rest of this section, we assume their necessary and sufficient conditions do not hold. It is known that these conditions do not hold for the normal distribution, in contrast to classes of stopping rules for the Poisson and binomial distributions, for example.

Assume now that a function $g(K, N)$ exists such that its expectation is zero for all values of the parameter and further that integrands are not zero almost everywhere over their integration ranges. Such a function must satisfy:

$$\int g(k, n) f_n(k) F(k) dk + \int g(k, 2n) f_{2n}(k) dk - \int \int g(k, 2n) f_n(z) f_n(k-z) F(z) dk dz = 0. \quad (16)$$

Substituting the general exponential form (5) into (16), and using (6), leads to

$$\begin{aligned} & \int h_n(k) e^{\theta k} \cdot \int g(k, n) h_n(k) F(k) e^{\theta k} dk \\ &= \int g(k, 2n) h_{2n}(k) e^{\theta k} dk - \int g(k, 2n) \left[\int h_n(z) h_n(k-z) F(z) dz \right] e^{\theta k} dk. \end{aligned} \quad (17)$$

Because the left hand side of (17) is a convolution, and using the uniqueness of the Laplace transform, we find:

$$g(k, 2n) = - \frac{\int g(z, n) h_n(z) h_n(k-z) F(z) dz}{h_{2n}(k) - \int h_n(z) h_n(k-z) F(z) dz}. \quad (18)$$

Hence, when $g(k, n)$ is chosen arbitrarily, (18) prescribes the choice for $g(k, 2n)$ which leads to a counterexample to completeness, hence establishing incompleteness.

For the CRSS case, when $F(k) \equiv F$, a constant, and also choosing $g(k, n) = c$, a constant, it follows that

$$g(k, 2n) = \frac{-F}{1-F} \cdot c.$$

In the limiting case of a deterministic stopping rule, $F(z) = 1$ over the stopping region \mathcal{S} and 0 over its set complement \mathcal{C} . It then follows that (14)–(15) reduce to:

$$p(k, n) = f_n(k) \cdot I(k \in \mathcal{S}), \quad (19)$$

$$p(k, 2n) = \int_{\mathcal{C}} f_n(z) f_n(k-z) dz. \quad (20)$$

For the deterministic case, (18) becomes:

$$g(k, 2n) = - \frac{\int_{\mathcal{S}} g(z, n) h_n(z) h_n(k-z) dz}{h_{2n}(k) - \int_{\mathcal{S}} h_n(z) h_n(k-z) dz} = - \frac{\int_{\mathcal{S}} g(z, n) h_n(z) h_n(k-z) dz}{\int_{\mathcal{C}} h_n(z) h_n(k-z) dz}. \quad (21)$$

Expression (21) follows from the fact that, in the deterministic case, $F(k) = 1$ over the stopping region \mathcal{S} and 0 elsewhere. The transition from one denominator to the other follows from observing that the convolution of $f_n(k)$ with itself produces $f_{2n}(k)$, and then replacing all of these by their explicit exponential-family form (5). Alternatively, it is easy to show that (21) follows immediately from the definition of a function $G(K, N)$ and (19)–(20).

The implication of these findings is that whenever they hold, the Lehmann-Scheffé theorem cannot be applied (see Section 2). It follows that a best mean-unbiased estimator does not necessarily exist for the average. In the next section, it will be shown that this is indeed the case for many, but not all outcome distributions and stopping rules, given that, for example, the exponential distribution does admit a uniform optimum. It will be shown that no optimum exists for the normal case, in line with Molenberghs *et al* (2012), and neither for the Bernoulli and Poisson cases, for a wide class of stopping rules.

3.2 The Normal Case

Following Molenberghs *et al* (2012), consider the outcome to be standard normal with mean μ and let stopping be governed by (10). They derived from first principles that the marginal probability of stopping is:

$$P(N = n) = \Phi \left(\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}} \right). \quad (22)$$

This expression also follows as a special case of (9) by choosing (10) as the stopping rule, i.e., $\tilde{f}_1(z)$ as the standard normal density and $A(k) = \alpha + \beta k/n$, and further $f_{n,\theta=\mu} = \varphi_{\mu,n}(k)$, where $\varphi_{\mu,s}(k)$ is the normal density with mean μ and variance s . Details of this derivation are provided in Appendix A.

Clearly, (22) depends on μ , implying that this pattern-mixture formulation is non-separable. In contrast, although the observed data are present in the conditional stopping probability, μ is not, implying separability in the selection model formulation.

In this case (14)–(15) takes the form

$$p_\mu(N, k) = p_0(N, k) \cdot \exp\left(k\mu - \frac{1}{2}n\mu^2\right) \quad (23)$$

with

$$p_0(n, k) = \phi_n(k) \cdot \Phi\left(\alpha + \frac{\beta}{n}k\right), \quad (24)$$

$$p_0(2n, k) = \phi_{2n}(k) \cdot \left[1 - \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}}\right)\right]. \quad (25)$$

Here, $\phi_s(k)$ is the normal density with mean 0 and variance s . Expression (25) is more explicit than (15), making use of the fact that the outcome densities are normal and the stopping probability is written as a normal cumulative distribution function. The derivation can be found in Molenberghs *et al* (2012). Based on the fact that integrating the joint densities specified by (23)–(25) over K and summing over N should be equal to one, leads to the identity:

$$\int \varphi_{\mu, n}(k) \cdot \Phi\left(\alpha + \frac{\beta}{n}k\right) dk = \int \varphi_{\mu, 2n}(k) \cdot \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}}\right) dk. \quad (26)$$

In Section 4.1, (26) will be derived in general.

The specific form of condition (18) is:

$$g(k, 2n) \cdot p_0(2n, k) = - \int \phi_n(k - z) \cdot g(z, n) \cdot \phi_n(z) \cdot \Phi\left(\alpha + \frac{\beta}{n}z\right) dz. \quad (27)$$

In the CRSS case, (24)–(25) reduce to:

$$p_0(n, k) = \phi_n(k) \cdot \Phi, \quad (28)$$

$$p_0(2n, k) = \phi_{2n}(k) \cdot (1 - \Phi), \quad (29)$$

where $\Phi \equiv \Phi(\alpha)$. Then here, as in the general case, (27) simplifies and leads to an explicit solution for a number of cases, especially when $g(k, n)$ is chosen to be a constant.

In addition, for this case, other explicit examples can be constructed, even when $\beta \neq 0$.

We reproduce the two examples of Molenberghs *et al* (2012).

Example 1. For the first of two examples, choose

$$g(k, n) = \tilde{\lambda}, \quad (30)$$

an arbitrary constant. Then it immediately follows from (27) that

$$g(k, 2n) = -\tilde{\lambda} \cdot \frac{\Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}}\right)}{1 - \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}}\right)}. \quad (31)$$

When $\beta = 0$, then the right hand side of (31) is constant and we can set $\tilde{\lambda} = \lambda(1 - \Phi)$, leading to $g(k, n) = \lambda(1 - \Phi)$ and $g(k, 2n) = -\lambda\Phi$.

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

$$\hat{\mu} = \frac{1}{N}K. \quad (32)$$

We can then construct a class of estimators derived there from.

Applying this to our example and choosing (30) and (31) for the special case of $\beta = 0$ leads to the following class of estimators:

$$\hat{\mu}_\lambda = \bar{\mu} + \lambda \cdot [(1 - \Phi)I(N = n) - \Phi I(N = 2n)]. \quad (33)$$

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

For the variance of (33), we obtain $\text{var}(\hat{\mu}_\lambda) = \text{var}(\bar{\mu}) + \lambda^2\Phi(1 - \Phi)$ 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 (32) or (51). before moving on to this, we first complete the second example.

When $\beta \rightarrow +\infty$, (24)–(25) reduces to

$$p_0(n, k) = \phi_n(k) \cdot I(k > 0), \quad (34)$$

$$p_0(2n, k) = \int_{k=-\infty}^{k=0} f_0(z, n) \cdot \phi_n(k - z) dz, . \quad (35)$$

In this case, the $G(K, N)$ functions will take a particular form.

With $g(k, n)$ as in (30), when $\beta \rightarrow +\infty$ (31) becomes

$$g(k, 2n) = -\tilde{\lambda} \cdot \frac{\Phi [(2n)^{-1/2}k]}{1 - \Phi [(2n)^{-1/2}k]}. \quad (36)$$

To see that the considerations particular to the above example are not unique, we consider a second one.

Example 2. Choose

$$g(k, n) = \frac{\lambda}{\Phi \left(\alpha + \frac{\beta}{n}k \right)}, \quad (37)$$

with λ a given constant, then

$$g(k, 2n) = -\frac{\lambda}{1 - \Phi \left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}} \right)}. \quad (38)$$

Choosing (37) and (38) now produces the estimator

$$\tilde{\mu}_\lambda = \bar{\mu} + \lambda \cdot \left[\frac{I(N = n)}{\Phi \left(\alpha + \frac{\beta}{n}k \right)} - \frac{I(n = 2n)}{1 - \Phi \left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}} \right)} \right]. \quad (39)$$

When now taking the limit $\beta \rightarrow +\infty$, (37)–(38) become:

$$g(k, n) = \lambda, \quad \text{defined for } k > 0, \quad (40)$$

$$g(k, 2n) = -\frac{\lambda}{1 - \Phi [(2n)^{-1/2}k]}. \quad (41)$$

The fact that the function $g(k, n)$ in (40) is undefined over the negative real numbers is unproblematic, because the stopping region is confined to the non-negative half line.

3.3 The Binary Case

While the binary case follows from the general considerations given in Section 3.1, it is insightful to examine this outcome type in some detail; here, integration is replaced by summation. Let the Bernoulli probability be π . The sample sum K then follows a $\text{Bin}(\pi, N)$ distribution and

$$f_{N,\pi}(k) = \binom{N}{k} \pi^k (1 - \pi)^{N-k}. \quad (42)$$

For now, as in the general case, we leave $F(k)$ unspecified. The joint distribution of (K, N) now takes the form

$$p(k, n) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} F(k), \quad (43)$$

$$p(k, 2n) = \pi^k (1 - \pi)^{2n-k} \left[\binom{2n}{k} - H(k) \right] = \pi^k (1 - \pi)^{2n-k} \tilde{H}(k), \quad (44)$$

where

$$H(k) = \sum_{z=0 \vee (k-n)}^{k \wedge n} \binom{n}{z} \binom{n}{k-z} F(z), \quad (45)$$

the meaning of $\tilde{H}(k)$ is obvious, $a \vee b = \max(a, b)$, and $a \wedge b = \min(a, b)$.

When stopping rule (13) is chosen, (43) becomes:

$$p(k, n) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} (p + q + 1) \binom{p+q}{p} \sum_{\ell=0}^q \frac{(-1)^\ell}{p + \ell + 1} \binom{q}{\ell} A(k)^{p+\ell+1}. \quad (46)$$

The marginal stopping probability can be derived by summing (46) over k but is generally unwieldy. In the particular case that $p = q = 0$ and $A(k) = k/n$, we find

$$p(k, n) = \frac{k}{n} \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \quad (47)$$

$$p(k, 2n) = \frac{2n-k}{2n} \binom{2n}{k} \pi^k (1 - \pi)^{2n-k}. \quad (48)$$

While the derivation of (47) is obvious, that of (48) is less straightforward and details are given in Appendix B. From (47), we deduce immediately that

$$P(N = n) = \frac{1}{n} \sum_{k=0}^n k \binom{n}{k} \pi^k (1 - \pi)^{n-k} = \pi.$$

In other words, this particular choice of conditional stopping rule produces essentially the simplest possible marginal stopping probability that depends on the parameter π that governs the outcomes.

The condition for the existence of a non-trivial function $g(K, N)$ with expectation zero for all π is a discrete version of (16) and reads:

$$\sum_{k=0}^n g(k, n) F(k) \binom{n}{k} \pi^k (1 - \pi)^{n-k} + \sum_{k=0}^{2n} g(k, 2n) \tilde{H}(k) \pi^k (1 - \pi)^{2n-k} = 0. \quad (49)$$

Writing $\gamma = \pi/(1 - \pi)$, (49) becomes

$$\sum_{k=0}^n g(k, n) F(k) \binom{n}{k} \gamma^k + \sum_{k=0}^{2n} g(k, 2n) \tilde{H}(k) (1 - \pi)^n \gamma^k = 0. \quad (50)$$

Using the discrete-data version of (6), i.e.,

$$(1 - \pi)^{-n} = \sum_{k=0}^n \binom{n}{k} \gamma^k,$$

it follows that

$$-\sum_{k=0}^{2n} \left[\sum_{z=0 \vee (k-n)}^{k \wedge n} \binom{n}{z} \binom{n}{k-z} g(z, n) F(z) \right] \theta^k = \sum_{k=0}^{2n} g(k, 2n) \tilde{H}(k) \theta^k.$$

Owing to equality of polynomial coefficients, we find:

$$g(k, 2n) = - \frac{\sum_{z=0 \vee (k-n)}^{k \wedge n} \binom{n}{z} \binom{n}{k-z} g(z, n) F(z)}{\tilde{H}(k)},$$

the discrete-data version of (18).

4 Generalized Sample Averages

4.1 The General Case

To underscore the impact of incompleteness of the statistics (K, N) , Molenberghs *et al* (2012) generalized the sample average (32) to

$$\bar{\mu} = \frac{K}{N} \cdot [c \cdot I(N = n) + d \cdot I(N = 2n)] = K \cdot \left[\frac{c \cdot I(N = n)}{n} + \frac{d \cdot I(N = 2n)}{2n} \right], \quad (51)$$

for some constants c and d . We will refer to it as the *generalized sample average* (GSA). The ordinary sample average follows as $c = d = 1$. In this section, (51) will be considered from a general exponential-family perspective. Sections 4.2 and 4.3 bring out some further specifics for the normal and Bernoulli cases, respectively.

From (5), the mean follows as $\mu = \partial a(\theta)/\partial \theta$. The expectation is:

$$E(\bar{\mu}) = \frac{c}{n} \int k f_n(k) F(k) dk + \frac{d}{2n} \int k f_{2n}(k) dk - \frac{d}{2n} \iint k f_n(z) f_n(k-z) F(z) dk dz. \quad (52)$$

This form can be simplified. We will derive two identities that are useful here and in what follows. Because integrating (14)–(15) over K and summing over N should lead to unity, it follows that

$$\int f_n(k) F(k) dk = \iint f_n(z) f_n(k-z) F(z) dk dz. \quad (53)$$

This equation obviously also follows from first principles. Likewise, we have that

$$\begin{aligned} \iint k f_n(z) f_n(k-z) F(z) dk dz &= \int_z f_n(z) F(z) \left[\int_k k f_n(k-z) dk \right] dz \\ &= \int_z f_n(z) F(z) [n\mu + z] dz \\ &= n\mu A_n(\mu) + B_n(\mu), \end{aligned} \quad (54)$$

where

$$A_n(\mu) = \int f_n(k) F(k) dk, \quad B_n(\mu) = \int k f_n(k) F(k) dk. \quad (55)$$

Using (54), we can rewrite (52) as

$$E(\bar{\mu}) = d \cdot \mu + \frac{2c-d}{2n} B_n(\mu) - \frac{d \cdot \mu}{2} A_n(\mu) \quad (56)$$

$$= d \cdot \mu + \frac{1}{2} \left[(2c-d) E \left(\frac{K}{N} \middle| N=n \right) - d \cdot \mu \right] \cdot P(N=n). \quad (57)$$

While obvious, it is useful to spell out (56)–(57) for the ordinary sample average:

$$E(\bar{\mu}) = \mu + \frac{1}{2n} B_n(\mu) - \frac{\mu}{2} A_n(\mu) \quad (58)$$

$$= \mu + \frac{1}{2} \left[E \left(\frac{K}{N} \middle| N=n \right) - \mu \right] \cdot P(N=n). \quad (59)$$

It is very intuitive that the bias in the sample average is a simple function of the difference between conditional and marginal expectation of K/N on the one hand, and the probability of stopping on the other.

The specific form of (52) will depend on both the exponential family member considered and the form of the stopping rule. In general, the expectation may be a non-linear function of μ and hence there may be no constants c and d for which the expectation is μ . Hence, in many situations, all linear estimators of the form (51) may be biased. Examples are given in Sections 4.2 and 4.3.

We now turn to the asymptotic behavior of the GSA, i.e., the case where $n \rightarrow +\infty$. Because K converges to a $N(n\mu, n\sigma^2)$ variable, and using a first-order Taylor series expansion $F(k) \approx F(n\mu) + F'(n\mu)(k - n\mu)$, we find from first principles:

$$A_n(\mu) \approx F(n\mu), \quad (60)$$

$$B_n(\mu) \approx n\mu F(n\mu) + n\sigma^2 F'(n\mu). \quad (61)$$

Using (60) and (61), (56) converges to:

$$E(\bar{\mu}) \xrightarrow{n \rightarrow +\infty} d \cdot \mu + (c - d) \cdot \mu \lim_{n \rightarrow +\infty} F(n\mu) + \frac{2c - d}{2} \sigma^2 \lim_{n \rightarrow +\infty} F'(n\mu). \quad (62)$$

In particular, for the ordinary sample average:

$$E(\bar{\mu}) \xrightarrow{n \rightarrow +\infty} \mu + \frac{1}{2} \sigma^2 \lim_{n \rightarrow +\infty} F'(n\mu). \quad (63)$$

In Section 4.2, we we will see that (62) is finite and, moreover, (63) equals μ . Sufficient conditions for this to hold in general can be given. Assume that $F(\cdot)$ is a continuously differentiable function that depends on k as a function of k/n . To emphasize this, write

$$F(k) = F(\eta(k/n)). \quad (64)$$

Then $F(n\mu) = F(\eta(\mu))$, independent of n and $F'(n\mu) = \eta'(\mu)F'(\eta(\mu))/n$, which depends on n only through the factor n^{-1} and hence converges to zero. More generally, a stop-

ping rule that satisfies $F'(n\mu) \xrightarrow{n \rightarrow +\infty} 0$ ensures that the sample average is asymptotically unbiased.

For a GSA to be asymptotically unbiased, (62) should equal μ . Assume that the third term on the right hand side of (62) is zero and $F(n\mu) = \tilde{F}(\mu)$ does not depend on n . The GSA is unbiased if $d + (c - d)\tilde{F}(\mu) = 1$ for all values of μ (note that, when $\mu = 0$, the limit is trivially equal to zero). This equation can be satisfied if $\tilde{F}(\mu)$ is constant, i.e., in the CRSS case to be discussed next. Otherwise, the equation can be satisfied only for $c = d = 1$, i.e., the ordinary sample average.

For the GSA to be unbiased in the finite-sample case, (56) needs to equal μ , leading to the requirement:

$$d = \frac{2\mu - 2c\tilde{\mu}P(N = n)}{2\mu - (\tilde{\mu} + \mu)P(N = n)}, \quad (65)$$

with $\tilde{\mu} = E(K/N|N = n)$. Evidently, this is a function of μ in the non-CRSS case and hence no uniformly unbiased estimator exists. Further, unless in the CRSS case, the ordinary sample average never satisfies (65) because this would imply that $\tilde{\mu} = \mu$ and hence the stopping probability would be independent of μ .

In the specific case of a CRSS, the constant F is taken out of the integrals on the right hand side of (56) and we easily find:

$$E_{\text{CRSS}}(\bar{\mu}) = [cF + d(1 - F)]\mu, \quad (66)$$

which is unbiased if and only if

$$d = \frac{1 - cF}{1 - F}. \quad (67)$$

An obvious solution is $c = d = 1$, the sample average, next to an infinite number of unbiased linear estimators of the type (51). Note that (67) follows from (65) upon observing that in the CRSS case $\mu = \tilde{\mu}$ and $P(N = n) = F$.

In addition to studying the overall expectation of the GSA, it is of interest to consider

the conditional expectations. These are:

$$E(\bar{\mu}|N = n) = \frac{c}{n} \cdot \frac{B_n(\mu)}{A_n(\mu)}, \quad (68)$$

$$E(\bar{\mu}|N = 2n) = \frac{d}{2n} \cdot \frac{2n\mu - n\mu A_n(\mu) - B_n(\mu)}{1 - A_n(\mu)}. \quad (69)$$

The ordinary sample average versions follow by setting $c = d = 1$ in (68)–(69).

The asymptotic behavior of (68)–(69), follows from applying (60) and (61):

$$E(\bar{\mu}|N = n) \xrightarrow{n \rightarrow +\infty} c \left(\mu + \sigma^2 \lim_{n \rightarrow +\infty} \frac{F'(n\mu)}{F(n\mu)} \right), \quad (70)$$

$$E(\bar{\mu}|N = 2n) \xrightarrow{n \rightarrow +\infty} d \left(\mu - \frac{1}{2}\sigma^2 \lim_{n \rightarrow +\infty} \frac{F'(n\mu)}{1 - F(n\mu)} \right). \quad (71)$$

For the ordinary sample average, when $F'(n\mu)$ converges to zero, the conditional expectations converge to μ . In case the limits in (70) and (71) differ from zero, there is a choice for c and d that produces conditional expectations equal to μ : $c = \mu/[\mu + \sigma^2 Q_1(\mu)]$ and $d = \mu/[\mu - 0.5\sigma^2 Q_2(\mu)]$, with obvious notation. Evidently, these are not uniform and therefore not useful in practice. These values for c and d lie at different sides of unity. We will return to the implications of limiting expressions (63) and (70)–(71) in Section 4.2.

A natural follow-up question is whether there is a, perhaps a uniform, optimal estimator in the CRSS case. From straightforward algebra we find that

$$\text{var}(\bar{\mu}) = \mu^2(1 - c)^2 \left(\frac{F}{1 - F} \right) + \frac{\sigma^2}{n} \frac{1 - 2Fc + 2Fc^2 - F^2c^2}{2(1 - F)}, \quad (72)$$

which is minimal for

$$c_{\text{opt}} = \frac{2\mu^2n + \sigma^2}{2\mu^2n + \sigma^2(2 - F)}, \quad d_{\text{opt}} = \frac{2\mu^2n + 2\sigma^2}{2\mu^2n + \sigma^2(2 - F)}. \quad (73)$$

In (72) and (73), σ^2 is the variance. It follows as either the first derivative of the mean function or, in the slightly more general case where there is an overdispersion parameter, as the first derivative of the mean multiplied with the overdispersion parameter.

Table 1: *Coefficients for optimum unbiased generalized sample average estimators, in the case of a completely random sample size.*

Exp. fam. member	c	d
Normal	$\frac{2\mu^2n + \sigma^2}{2\mu^2n + \sigma^2(2 - F)}$	$\frac{2\mu^2n + 2\sigma^2}{2\mu^2n + \sigma^2(2 - F)}$
Bernoulli	$\frac{2\pi n + (1 - \pi)}{2\pi n + (1 - \pi)(2 - F)}$	$\frac{2\pi n + 2(1 - \pi)}{2\pi n + (1 - \pi)(2 - F)}$
Poisson	$\frac{2\lambda n + 1}{2\lambda n + 2 - F}$	$\frac{2\lambda n + 2}{2\lambda n + 2 - F}$
Exponential	$\frac{2n + 1}{2n + 2 - F}$	$\frac{2n + 2}{2n + 2 - F}$

Whereas constraint (67) on the pair (c, d) does not depend on the particular exponential family considered, rather only on the constant probability of stopping, this is not true for the optimality condition (73). Because of its dependence on μ and σ^2 , (73) will not generally allow for a uniform optimum, except in specific examples. A few examples are given in Table 1. As Molenberghs *et al* (2012) observed for the normal case, most solutions indeed indicate that there is no uniform minimum, even though all coefficients converge to 1 if the sample size increases. A noteworthy exception is the exponential family distribution, for which there is a uniform solution common to all values of the mean parameter and different from 1, for every value of the sample size n .

In all cases, when $F = 0$ then $d = 1$ and c is irrelevant, while for $F = 1$, the reverse is true.

We have seen above that, even for CRSS, the sample average is not optimal, and that there is no uniform optimal solution, even though the sample average approximately is. The exponential case is an exception to this, as we saw above. 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.

The asymptotic behavior for a deterministic stopping rule is completely captured by the normal case, described in Section 4.2, because the stopping rule $F(k)$ has the effect of restricting the integrals over the stopping and continuation regions \mathcal{S} and \mathcal{C} , respectively. This, together with the fact that $f_n(k)$ approaches a normal density with mean $n\mu$ and variance $n\sigma^2$ establishes this fact. As a result, we can restrict considerations regarding the deterministic case to the finite-sample situation. But also this one is very straightforward. Given that the joint distribution (14)–(15) becomes (19)–(20), the functions $A_n(\mu)$ and $B_n(\mu)$ in (55) take the form:

$$A_n(\mu) = \int_{\mathcal{S}} f_n(k) dk, \quad B_n(\mu) = \int_{\mathcal{S}} k f_n(k) dk. \quad (74)$$

and all results, such as marginal and conditional expectations of the GSA, carry over.

4.2 The Normal Case

Molenberghs *et al* (2012) showed that expectation (52) of generalized sample average (51) becomes, for the normal case with probit stopping probability:

$$E(\bar{\mu}) = d\mu + (c - d)\mu\Phi(\nu) + \frac{2c - d}{2n} \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \phi(\nu) \quad (75)$$

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

The specific case of a CRSS, here corresponding with $\beta = 0$, has been considered in Section 4.1.

When $\beta \neq 0$, expression (75) does not in general simplify. It is easy to see here that there cannot be a uniformly unbiased estimator, i.e., that there cannot exist c and d such that (75) reduces to μ , for all μ , and in particular for $\mu = 0$. For this special case

$$0 = \frac{2c - d}{2n} \cdot \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \phi(\nu_0),$$

where $\nu_0 = (\alpha)/\sqrt{1 + \beta^2/n}$. Given that $\beta \neq 0$, this expression leads to the condition $2c = d$. Substituting this back into (75), which should be μ for every value of μ , and not just for $\mu = 0$, produces $E(\bar{\mu}) = c\mu[2 - \Phi(\nu)]$, which equals μ only if $c = [2 - \Phi(\nu)]^{-2}$. Based on this, 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 $\beta = 0$.

Molenberghs *et al* (2012) quantified the asymptotic bias. In Section 4.1 this was done in general for CRSS. Turning to the case of $\beta \neq 0$, Molenberghs *et al* (2012) began with the ordinary sample average $c = d = 1$, which leads to expectation:

$$E(\hat{\mu}) = \mu + \frac{1}{2n} \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \phi(\nu) \xrightarrow{n \rightarrow +\infty} \mu + \frac{1}{2n} \beta \cdot \phi(\alpha + \beta\mu) \xrightarrow{n \rightarrow +\infty} \mu. \quad (76)$$

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

$$E(\hat{\mu}) = \mu + \frac{1}{2\sqrt{n}} \cdot \phi(\sqrt{n}\mu) \xrightarrow{n \rightarrow +\infty} \mu. \quad (77)$$

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

$$E(\bar{\mu}) \xrightarrow{n \rightarrow +\infty} d\mu + (c - d)\mu\Phi(\alpha + \beta\mu), \quad (78)$$

which equals μ if and only if:

$$d = \frac{1 - c\Phi(\alpha + \beta\mu)}{1 - \Phi(\alpha + \beta\mu)}. \quad (79)$$

While (79) and (67) 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 above limits also follow from (62) and (63), because now $\eta(k/n) = \alpha + \beta k/n$ and the derivative therefore is $F'(n\mu) = \phi(\alpha + \beta\mu) \cdot \beta/n$, which leads to (76).

Molenberghs *et al* (2012) also studied the deterministic stopping rule case, following from $\beta \rightarrow \infty$, because then (78) becomes

$$E(\bar{\mu}) = d\mu + (c - d)\mu\Phi(\sqrt{n}\mu) + \frac{2c - d}{2\sqrt{n}}\phi(\sqrt{n}\mu) \xrightarrow{n \rightarrow +\infty} \begin{cases} c\mu & \text{if } \mu > 0, \\ d\mu & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0. \end{cases} \quad (80)$$

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 $c = d = 1$.

The pleasing asymptotic behavior of the sample average is connected to the choice of the stopping rule, in view of limiting expressions (63), (70), and (71). In this case, $\lim_{n \rightarrow +\infty} F(n\mu) = \Phi(\alpha + \beta\mu)$, a constant in $]0, 1[$, while $\lim_{n \rightarrow +\infty} F'(n\mu) = (\beta/n)\phi(\alpha + \beta\mu) = 0$. Hence, the limits of $F'(n\mu)$, $F'(n\mu)/F(n\mu)$, and $F'(n\mu)/[1 - F(n\mu)]$ are zero. The essence is that the stopping rule is a cumulative density function based transformation of a linear predictor in k/n . It is therefore of interest to examine the consequences of switching to a different class of stopping rule. Therefore, we change the stopping rule to $\Phi(\alpha + \beta k)$. Then $F'(n\mu) = \beta\phi(\alpha + \beta n\mu)$ which again tends to zero. However, depending on the sign of β and μ , $\Phi(\alpha + \beta n\mu)$ tends to either zero or one. Applying de l'Hôpital's rule to the case where $F(n\mu)$ tends to zero as well, produces $-\beta(\alpha + \beta n\mu)$ which tends to infinity, and hence the regularity condition (70) appears *not* to be satisfied. This requires careful qualification, because not only does $F(n\mu)$ appear in (70), it is also the probability with which $N = n$, which then equally well tends to zero. Thus, for this case, in the limit, $E(\bar{\mu}|N = 2n) = E(\bar{\mu})$ and unbiasedness still applies. Evidently, when $1 - F(n\mu)$ tends to zero rather than $F(n\mu)$, we are in the mirror image of the above situation, and the result is the same. This result applies more generally. If $F(k) = \Phi(\alpha + \beta kn^m)$, with m any real number, then $F'(n\mu)$ converges to zero whatever m is. Further, $F(n\mu)$ converges

to $\Phi(\alpha + \beta\mu)$ for $m = -1$, $\Phi(\alpha)$ for $m < -1$, and $\Phi(\pm\infty)$ (i.e., 0 or 1) for $m > -1$. This means that the sample average is asymptotically unbiased in all cases, and even conditionally asymptotically unbiased, based on the same logic as before.

4.3 The Binary Case

An explicit form for the expectation of the generalized sample average in the Bernoulli case is

$$E(\bar{\pi}) = d\pi + \frac{c}{n} \sum_{k=0}^n k \binom{n}{k} \pi^k (1-\pi)^{n-k} F(k) - \frac{d}{2n} \sum_{k=0}^{2n} k \pi^k (1-\pi)^{2n-k} H(k), \quad (81)$$

with $H(k)$ as in (45).

The CRSS has been covered in Section 4.1, and the coefficients for optimal estimators listed in Table 1.

As an example, when stopping rule (13) is chosen, with $p = q = 0$ and $A(k) = k/n$, we have that $F(k) = A(k) = k/n$ and

$$H(k) = k/2n \cdot \binom{2n}{k}.$$

Hence, (81) becomes

$$\begin{aligned} E(\bar{\pi}) &= d\pi + \frac{c}{n^2} E_n(K^2) - \frac{d}{(2n)^2} E_{2n}(K^2) \\ &= \pi \left[d + \frac{c}{n}(1 - \pi + n\pi) - \frac{d}{2n}(1 - \pi + 2n\pi) \right]. \end{aligned} \quad (82)$$

Clearly, the estimator is unbiased if and only if

$$d = \frac{1 - \frac{c}{n}(1 - \pi + n\pi)}{1 - \frac{1}{2n}(1 - \pi + 2n\pi)}.$$

Hence, there is no uniform solution, neither in π nor in n . When $n \rightarrow +\infty$,

$$d \xrightarrow{n \rightarrow +\infty} \frac{1 - c\pi}{1 - \pi}. \quad (83)$$

Note that the ordinary sample average, i.e., $c = d = 1$, is a solution to (83), as it should.

Turning to the case of a deterministic stopping rule, assume that the stopping region \mathcal{S} is defined by $(k \leq k_0)$, i.e., $F(k) = 1$ if $k \leq k_0$ and 0 otherwise. Functions $A_n(\pi)$ and $B_n(\pi)$ as in (74) are here:

$$A_n(\pi) = \sum_{k=0}^{k_0} \binom{n}{k} \pi^k (1 - \pi)^{n-k} = \mathcal{I}(k_0, n, \pi), \quad (84)$$

$$B_n(\pi) = \sum_{k=0}^{k_0} k \binom{n}{k} \pi^k (1 - \pi)^{n-k} = n\pi \mathcal{I}(k_0 - 1, n - 1, \pi). \quad (85)$$

$\mathcal{I}(k, n, \pi)$, the binomial cumulative distribution function, is actually defined by (84). Various alternative formulations exist, but none is of direct use to us here. The expectation of the GSA becomes:

$$E(\bar{\pi}) = \pi \left[d + \frac{2c - d}{2} \mathcal{I}(k_0 - 1, n - 1, \pi) - \frac{d}{2} \mathcal{I}(k_0, n, \pi) \right]. \quad (86)$$

For the ordinary sample average, (86) reduces to

$$E(\bar{\pi}) = \pi \left\{ 1 + \frac{1}{2} [\mathcal{I}(k_0 - 1, n - 1, \pi) - \mathcal{I}(k_0, n, \pi)] \right\}.$$

5 Likelihood Estimators

5.1 The General Case

For notational convenience, we introduce the indicator variable $Z = I(N = n)$.

The joint likelihood for the observed data and stopping occurrence is:

$$L(\mu) = h_N(k) \exp\{\theta k - Na(\theta)\} \cdot F(k)^z \cdot [1 - F(k)]^{1-z}. \quad (87)$$

Likelihood decomposition (87) is of a selection model type. The factors pertaining to stopping are free of the mean parameter μ . This simplifies the kernel of the likelihood

$\ell(\mu)$, score function $S(\mu)$, and Hessian $H(\mu)$:

$$\ell(\mu) = \ln h_N(k) + \theta k - Na(\theta), \quad (88)$$

$$S(\mu) = k - Na'(\theta) = k - N\mu, \quad (89)$$

$$H(\mu) = -Na''(\theta) = -N\mu'. \quad (90)$$

The simplicity of this estimator is a direct consequence of ignorability. Based on (14)–(15), the conditional probability for the sample sum K , given the sample size N , can be derived. For the case that $N = n$, the likelihood function is:

$$L_n(\mu) = \frac{F(k)h_n(k)e^{\theta k}}{\int F(k)h_n(k)e^{\theta k}dk}, \quad (91)$$

leading to the following expressions for the log-likelihood, score, and Hessian:

$$\ell_n(\mu) = \ln F(k) + \ln h_n(k) + \theta k - \ln \int F(k)h_n(k)e^{\theta k}dk, \quad (92)$$

$$S_n(\mu) = k - \frac{B_n(\mu)}{A_n(\mu)} = k - E(K|N = n), \quad (93)$$

$$\begin{aligned} H_n(\mu) &= - \left[\frac{C_n(\mu)}{A_n(\mu)} - \left\{ \frac{B_n(\mu)}{A_n(\mu)} \right\}^2 \right] \\ &= - [E(K^2|N = n) - E(K|N = n)^2] = -\text{var}(K|N = n). \end{aligned} \quad (94)$$

Here $A_n(\mu)$ and $B_n(\mu)$ are as defined in (55), and

$$C_n(\mu) = \int k^2 f_n(k)F(k)dk.$$

When $N = 2n$, the likelihood takes the form:

$$L_{2n}(\mu) = \frac{D_n(\mu)}{1 - A_n(\mu)}, \quad (95)$$

with

$$D_n(\mu) = \exp\{\theta k - 2na(\theta)\} \left[h_{2n}(k) - \int h_n(z)h_n(k - z)F(z)dz \right].$$

Then, the counterparts to (92)–(94) are: Then, the counterparts to (92)–(94) are:

$$\begin{aligned} \ell_{2n}(\mu) &= \theta k - 2na(\theta) + \ln \left[h_{2n}(k) - \int h_n(z)h_n(k-z)F(z)dz \right] \\ &\quad - \ln \{1 - A_n(\mu)\}, \end{aligned} \quad (96)$$

$$\begin{aligned} S_{2n}(\mu) &= k - \frac{2n\mu - n\mu A_n(\mu) - B_n(\mu)}{1 - A_n(\mu)} \\ &= k - E(K|N = 2n), \end{aligned} \quad (97)$$

$$\begin{aligned} H_{2n}(\mu) &= -\frac{2n\sigma^2 - n\sigma^2 A_n(\mu) + (n\mu)^2 A_n(\mu) - C_n(\mu)}{1 - A_n(\mu)} \\ &\quad + \frac{[B_n(\mu) - n\mu A_n(\mu)][2n\mu - n\mu A_n(\mu) - B_n(\mu)]}{[1 - A_n(\mu)]^2} \\ &= -[E(K^2|N = 2n) - E(K|N = 2n)^2] = -\text{var}(K|N = 2n). \end{aligned} \quad (98)$$

From the form of (93) and (97), it is immediately clear that the conditional expectations of the conditional scores are equal to zero and therefore also the marginal expectation.

The expectation of the joint likelihood based estimator, which is the ordinary sample average, was presented in Section 4.1. Even though there is small-sample bias in most cases different from CRSS, wide classes of stopping rules are asymptotically unbiased. The bias expressions in the conditional expectation of the sample average, which of course are also the bias expressions for the joint likelihood estimator, are of the form $E(K/N|N) - \mu$. These expressions coincide with the correction in conditional score equations (93) and (97) relative to (89), which follows immediately upon rewriting the former as $S_N(\mu) = k - N\mu + \{N\mu - E(K|N)\}$.

Turning to precision and information, first note that for CRSS, $H_n(\mu) = -n\sigma^2$ and $H_{2n}(\mu) = -2n\sigma^2$; hence the marginal and conditional information in this case reduces to $I(\mu) = I_c(\mu) = n\sigma^2(2 - F)$.

In the general case, the marginal and conditional information are

$$I(\mu) = n\sigma^2[2 - A_n(\mu)], \quad (99)$$

$$I_c(\mu) = n\sigma^2[2 - A_n(\mu)] - \frac{[n\mu A_n(\mu) - B_n(\mu)]^2}{A_n(\mu)[1 - A_n(\mu)]}. \quad (100)$$

Using information expressions (99)–(100), the bias for the marginal likelihood estimator, and the fact that the conditional likelihood estimator is unbiased, the mean squared error expressions are:

$$\text{MSE}_n(\hat{\mu}) = \frac{1}{n\sigma^2[2 - A_n(\mu)]} + \frac{1}{4n^2} [n\mu A_n(\mu) - B_n(\mu)]^2, \quad (101)$$

$$\begin{aligned} \text{MSE}_n(\hat{\mu}_c) &= \frac{1}{n\sigma^2[2 - A_n(\mu)]} \\ &+ \frac{[n\mu A_n(\mu) - B_n(\mu)]^2}{A_n(\mu)[1 - A_n(\mu)]\{n\sigma^2[2 - A_n(\mu)]\}^2 - n\sigma^2[2 - A_n(\mu)][n\mu A_n(\mu) - B_n(\mu)]^2}. \end{aligned} \quad (102)$$

Recall that for CRSS $B_n(\mu) = n\mu A_n(\mu)$ and both MSE expressions coincide. In the asymptotic case, (101)–(102) can be approximated, using (60)–(61), as:

$$\text{MSE}_{n \rightarrow +\infty}(\hat{\mu}) \simeq \frac{1}{n\sigma^2[2 - F(n\mu)]} + \frac{\sigma^4}{4} F'(n\mu)^2, \quad (103)$$

$$\begin{aligned} \text{MSE}_{n \rightarrow +\infty}(\hat{\mu}_c) &\simeq \frac{1}{n\sigma^2[2 - F(n\mu)]} \\ &+ \frac{F'(n\mu)^2}{[2 - F(n\mu)] \{F(n\mu)[1 - F(n\mu)][2 - F(n\mu)] - n\sigma^2 F'(n\mu)^2\}} \end{aligned} \quad (104)$$

Returning to the exact expressions (101)–(102), it is relatively straightforward to show that (101) is smaller than (102) if and only if $\sigma^2 A_n(\mu)[1 - A_n(\mu)][2 - A_n(\mu)] \geq 4$. Requiring that this inequality is satisfied for all values of $A_n(\mu)$ in the unit interval comes down to requiring that $\sigma^2 \leq 2.54$. Hence, the MSE is smaller in the marginal case if the variance is sufficiently small. For binary data, for example this is always satisfied given that the variance takes the form $\pi(1 - \pi)$. Also, asymptotically, $A_n(\mu)$ typically tends to either 0 or 1, and the above requirement is then also satisfied. In case $F'(n\mu)$ tends to zero as n tends to infinity, both MSE expressions tend to the same limit.

5.2 The Normal Case

Molenberghs *et al* (2012) studied this case in detail. Because of the relatively simple expressions for the normal density and the probit stopping rule (22), additional insight can be gained. We summarize their arguments.

Joint-likelihood expressions (87)–(90) for this case are:

$$L(\mu) = \prod_{i=1}^N \phi(y_i; \mu) \cdot F\left(\alpha + \frac{\beta}{n}k\right)^z \cdot \left\{1 - F\left(\alpha + \frac{\beta}{n}k\right)\right\}^{1-z}, \quad (105)$$

$$\ell(\mu) = -\frac{1}{2} \sum_{i=1}^N (y_i - \mu)^2, \quad (106)$$

$$S(\mu) = \sum_{i=1}^N (y_i - \mu), \quad (107)$$

$$H(\mu) = -N. \quad (108)$$

By contrast, one can start from the conditional probability for the outcomes:

$$f(y_1, \dots, y_N | Z = z) = \frac{\prod_{i=1}^N \phi(y_i; \mu) \Phi\left(\alpha + \frac{\beta}{n}k\right)^z [1 - \Phi\left(\alpha + \frac{\beta}{n}k\right)]^{1-z}}{\Phi\left(\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}}\right)^z \left[1 - \Phi\left(\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}}\right)\right]^{1-z}}. \quad (109)$$

Write, for convenience, $\tilde{\alpha} = \alpha/\sqrt{1 + \beta^2/n}$ and $\tilde{\beta} = \beta/\sqrt{1 + \beta^2/n}$. Further, let $\nu = \tilde{\alpha} + \tilde{\beta}\mu$. Consider first the case where $N = n$. The kernel of the likelihood $\ell(\mu)$, score function $S(\mu)$, and Hessian $H(\mu)$ are:

$$\ell_n(\mu) = -\frac{1}{2} \sum_{i=1}^N (y_i - \mu)^2 - \ln \Phi(\nu), \quad (110)$$

$$S_n(\mu) = \sum_{i=1}^N (y_i - \mu) - \tilde{\beta} \cdot \frac{\phi(\nu)}{\Phi(\nu)}, \quad (111)$$

$$H_n(\mu) = -N + \tilde{\beta}^2 \cdot [\nu \cdot \Phi(\nu) + \phi(\nu)] \cdot \frac{\phi(\nu)}{\Phi(\nu)^2}. \quad (112)$$

When $N = 2n$, the corresponding expressions are:

$$\ell_{2n}(\mu) = -\frac{1}{2} \sum_{i=1}^N (y_i - \mu)^2 - \ln [1 - \Phi(\nu)], \quad (113)$$

$$S_{2n}(\mu) = \sum_{i=1}^N (y_i - \mu) + \tilde{\beta} \cdot \frac{\phi(\nu)}{1 - \Phi(\nu)}, \quad (114)$$

$$H_{2n}(\mu) = -N - \tilde{\beta}^2 \cdot \{\nu \cdot [1 - \Phi(\nu)] - \phi(\nu)\} \cdot \frac{\phi(\nu)}{[1 - \Phi(\nu)]^2}. \quad (115)$$

Next, we consider bias, consistency, precision, and mean squared error of the joint and conditional likelihood estimators.

In the CRSS case, μ vanishes from the joint stopping model, and both estimators coincide with the ordinary sample average, amply studied in Section 4.

Asymptotic unbiasedness of the sample average follows simultaneously from direct calculation as well as from the fact that it is the maximum likelihood estimator from the joint likelihood (106). In terms of the conditional likelihood, the estimator is obtained from the solution to the score equations, (111) and (114). These can be reformulated as:

$$\tilde{S}(\mu) = \frac{1}{N} \sum_{i=1}^N y_i - \mu - \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \phi(\nu) \cdot \left\{ \frac{I(N = n)}{n \cdot \Phi(\nu)} - \frac{I(N = 2n)}{2n \cdot [1 - \Phi(\nu)]} \right\}. \quad (116)$$

The expectation of (116) results from (76), combined with the observation that the probability of stopping is $\Phi(\nu)$:

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

Finite-sample unbiasedness follows directly from the linearity of the score in the data. Thus, 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.

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, in agreement with our results of Section 5.1. To see this explicitly, it is convenient to rewrite the expectation of the generalized sample average (75)

$$E(\bar{\mu}) = c \cdot \left\{ \mu + \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \frac{\phi(\nu)}{n\Phi(\nu)} \right\} \cdot \Phi(\nu) + d \cdot \left\{ \mu - \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \frac{\phi(\nu)}{2n[-1\Phi(\nu)]} \right\} \cdot [1 - \Phi(\nu)],$$

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

$$E(\bar{\mu}|N = n) = \mu + \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \frac{\phi(\nu)}{n\Phi(\nu)}, \quad (117)$$

$$E(\bar{\mu}|N = 2n) = \mu - \frac{\beta}{\sqrt{1 + \beta^2/n}} \cdot \frac{\phi(\nu)}{2n[1 - \Phi(\nu)]}. \quad (118)$$

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.

Turning to precision, the expected information in the joint approach is

$$I(\mu) = E(N) = n \cdot \Phi(\tilde{\alpha} + \tilde{\beta}\mu) + 2n \cdot [1 - \Phi(\tilde{\alpha} + \tilde{\beta}\mu)] = n[2 - \Phi(\tilde{\alpha} + \tilde{\beta}\mu)], \quad (119)$$

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

$$I_c(\mu) = n[2 - \Phi(\tilde{\alpha} + \tilde{\beta}\mu)] - \frac{\tilde{\beta}^2 \phi(\tilde{\alpha} + \tilde{\beta}\mu)^2}{\Phi(\tilde{\alpha} + \tilde{\beta}\mu)[1 - \Phi(\tilde{\alpha} + \tilde{\beta}\mu)]} \quad (120)$$

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

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

The difference between joint and conditional information tends to zero when n tends to infinity.

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 that conditioning is done on a non-ancillary statistic. In line with the general theory in Section 5.1, 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.

Because of the opposing results for bias and precision, it is useful to calculate the mean squared error for both estimators. The expressions from Molenberghs *et al* (2012), are

$$\text{MSE}(\hat{\mu}) = \frac{1}{n[2 - \Phi(\nu)]} + \frac{1}{4n^2} \tilde{\beta}^2 \phi(\nu)^2, \quad (121)$$

$$\text{MSE}(\hat{\mu}_c) \simeq \frac{1}{n[2 - \Phi(\nu)]} + \frac{1}{[2 - \Phi(\nu)]^2 \Phi(\nu) [1 - \Phi(\nu)] n^2} \tilde{\beta}^2 \phi(\nu)^2. \quad (122)$$

Comparing these, 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 $n \rightarrow \infty$, ν approaches $\alpha + \beta\mu$ and $\tilde{\beta}$ 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 (121) and (122) are of the order of $1/n^2$.

To conclude this section, we examine the above quantities for the limiting case of a deterministic stopping rule, i.e., $\beta \rightarrow \pm\infty$. Focusing on the positive limit, we obtain

$$\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}} \xrightarrow{\beta \rightarrow +\infty} \sqrt{n}\mu.$$

The marginal outcome model retains its normal-density form, while the other three expressions change. First, the conditional outcome model (109) becomes

$$f(y_1, \dots, y_N | Z = z) = \frac{\prod_{i=1}^N \phi(y_i; \mu)}{\Phi(\sqrt{n}\mu)^z [1 - \Phi(\sqrt{n}\mu)]^{1-z}}. \quad (123)$$

Second, (4) is given by $P(N = n | \mathbf{y}_i) = 1$ if $K > 0$ and 0 otherwise. Third, (22) takes the limiting form $P(N = n) = \Phi(\sqrt{n}\mu)$.

Molenberghs *et al* (2012) derived the following bias expressions under deterministic stopping:

$$E(\bar{\mu} | N = n) = \mu + \frac{\phi(\sqrt{n}\mu)}{\sqrt{n}\Phi(\sqrt{n}\mu)}, \quad (124)$$

$$E(\bar{\mu} | N = 2n) = \mu - \frac{\phi(\sqrt{n}\mu)}{2\sqrt{n}[1 - \Phi(\sqrt{n}\mu)]}, \quad (125)$$

from which it follows that $E[\tilde{S}(\mu)|N] = 0$. For the sample average a little more caution is required. From (117) and (118), it follows that $E(\bar{\mu}|N)$ converges to μ at a rate of n , because $\nu \rightarrow \alpha + \beta\mu$. The situation is more subtle when $\beta \rightarrow \infty$. To show this, we take the limit of (124) and (125) as $n \rightarrow \infty$. When $\mu < 0$, applying de l'Hôpital's rule whenever needed, the limits are $E(\bar{\mu}|N = n) \rightarrow 0$ and $E(\bar{\mu}|N = 2n) \rightarrow \mu$. Similarly, when $\mu > 0$, the corresponding expressions are $E(\bar{\mu}|N = n) \rightarrow \mu$ and $E(\bar{\mu}|N = 2n) \rightarrow \mu/2$. It follows that when $\mu = 0$, these are both equal to 0. Somewhat surprising, this shows there is *no* bias in the conditional means: when $n \rightarrow \infty$, 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.

For precision, the second term in (120) 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 $n \rightarrow \infty$. This has the interesting consequence that there is no difference in precision when $\beta = 0$ and $\beta \rightarrow \infty$, but that there is for finite non-zero β .

For the mean squared error, the argument differs from the one used in the stochastic stopping rule case, because now $\nu = \sqrt{n}\mu$ and $\tilde{\beta} = \sqrt{n}$, which leads to

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

When $n \rightarrow \infty$, both expressions converge to $1/(2n)$ if the trial continues and $1/n$ if the trial stops, and the difference between them disappears.

5.3 The Binary Case

Joint-likelihood expressions for the binary case, in the probability parameter π are:

$$L(\pi) = \binom{N}{k} \pi^k (1 - \pi)^{N-k} \cdot F(k)^z \cdot [1 - F(k)]^{1-z}, \quad (126)$$

$$\ell(\pi) \propto k \ln \pi + (N - k) \ln(1 - \pi), \quad (127)$$

$$S(\pi) = \frac{k}{\pi} - \frac{N - k}{1 - \pi}, \quad (128)$$

$$H(\pi) = \frac{-K + 2\pi K - \pi^2 N}{\pi^2(1 - \pi)^2}. \quad (129)$$

The expected Hessian, for fixed sample size, is well known to be $-N/[\pi(1 - \pi)]$. However, with our stopping rule $F(k) = k/n$, it can be shown to be

$$E[H(\pi)] = -\frac{(2 - \pi)n}{\pi(1 - \pi)}. \quad (130)$$

Likewise, given that the solution to $S(\pi)$ is the sample average, the bias is

$$\text{Bias}(\hat{\pi}) = \frac{\pi(1 - \pi)}{2n}. \quad (131)$$

We will return to this in what follows.

Turning to the conditional expressions, for $N = n$, (91)–(94) become:

$$L_n(\pi) = \frac{\binom{n}{k} \pi^k (1 - \pi)^{n-k} F(k)}{A_n(\pi)}, \quad (132)$$

with

$$A_n(\pi) = \sum_{\ell=0}^n \binom{n}{\ell} \pi^\ell (1 - \pi)^{n-\ell} F(\ell),$$

leading to:

$$\ell_n(\pi) \propto k \ln \pi + (n - k) \ln(1 - \pi) - \ln A_n(\pi), \quad (133)$$

$$S_n(\pi) = \frac{k}{\pi} - \frac{n - k}{1 - \pi} - \frac{n\pi A_n(\pi) - B_n(\mu)}{A_n(\pi)}. \quad (134)$$

For the case where $N = 2n$ we obtain:

$$L_{2n}(\pi) = \frac{\binom{2n}{k} \pi^k (1-\pi)^{2n-k} [1 - F(k)]}{1 - A_n(\pi)}, \quad (135)$$

$$\ell_{2n}(\pi) \propto k \ln \pi + (2n - k) \ln(1 - \pi) - \ln[1 - A_n(\pi)], \quad (136)$$

$$S_{2n}(\pi) = \frac{k}{\pi} - \frac{2n - k}{1 - \pi} + \frac{n\pi A_n(\pi) - B_n(\mu)}{1 - A_n(\pi)}. \quad (137)$$

The fact that $E[S_N(\pi)|N] = 0$ follows from the derivations in Section 5.1, as well as from first principles.

It is clear that the above expressions are slightly different than the general expressions (87)–(90), because π is not the natural parameter. This does not prohibit further derivations but makes them cumbersome from an algebraic standpoint. Therefore, we switch to the logit form, i.e., $\alpha = \ln[\pi/(1 - \pi)]$ will be used. Furthermore, we restrict attention to the particular stopping rule used in previous sections, $F(k) = k/n$. Then, (126)–(129) become:

$$L(\alpha) = \binom{N}{k} \frac{e^{\alpha k}}{(1 + e^\alpha)^N} \cdot \left(\frac{k}{n}\right)^z \cdot \left(\frac{n - k}{n}\right)^{1-z}, \quad (138)$$

$$\ell(\alpha) \propto \alpha k - N \ln(1 + e^\alpha), \quad (139)$$

$$S(\alpha) = k - N\pi, \quad (140)$$

$$H(\alpha) = -N\pi(1 - \pi). \quad (141)$$

The use of π on the right hand sides of (140) and (141) rather than α is for convenience only. The expected Hessian is straightforward to derive, given that $E(N) = n(2 - \pi)$:

$$E[H(\alpha)] = -n(2 - \pi)\pi(1 - \pi). \quad (142)$$

In fact, this calculation is considerably easier than the derivation of (130), even though they are equivalent. Indeed, (130) follows from (142) by applying the delta method. Because $\pi = \text{expit}(\alpha)$, the derivative is $\Delta = \partial\alpha/\partial\pi = [\pi(1 - \pi)]^{-1}$, and $E[H(\pi)] = \Delta^2 \cdot E[H(\alpha)]$, as it should.

The forms for (133)–(137), supplemented with the Hessians, are:

$$\ell_n(\alpha) \propto \alpha k - n \ln(1 + e^\alpha) - \alpha + \ln(1 + e^\alpha), \quad (143)$$

$$S_n(\alpha) = (k - 1) - (n - 1)\pi, \quad (144)$$

$$H_n(\alpha) = -(n - 1)\pi(1 - \pi), \quad (145)$$

$$\ell_{2n}(\alpha) \propto \alpha k - 2n \ln(1 + e^\alpha) + \ln(1 + e^\alpha), \quad (146)$$

$$S_{2n}(\alpha) = k - (2n - 1)\pi, \quad (147)$$

$$H_{2n}(\alpha) = -(2n - 1)\pi(1 - \pi). \quad (148)$$

Note that the conditional Hessians are in line with what one would expect from conditioning upon the sample size: one ‘degree of freedom’ is removed for mean parameter estimation. Such an operation though, is standard only when the sample size is fixed. The counterintuitive effect on the efficiency was seen in general in Section 5.1 and very explicitly for the normal data setting in Section 5.2. Straightforward algebra then establishes:

$$E[H_N(\alpha)] = -\pi(1 - \pi)[(2 - \pi)n - 1] = E[H(\alpha)] + \pi(1 - \pi). \quad (149)$$

Thus, the conditional information is expected to take one subject less into account than the marginal expectation, precisely the opposite of what one would expect in the fixed sample-size case. The bias in the estimators is easy to quantify, given that the estimators are $\hat{\pi} = k/N$ in the marginal case and $\hat{\pi}_c = (k-1)/(n-1)$ when $N = n$ and $\hat{\pi}_c = k/(2n-1)$ when $N = 2n$. The biases are $(n - k)/([n(n - 1)])$ and $-k/[2n(2n - 1)]$, respectively. This follows from the difference between the marginal and conditional estimators, given that the latter is unbiased. For this stopping rule, $E(K|N = n) = n\pi + 1 - \pi$ and $E(K|N = 2n) = \pi(2n - 1)$, and so the average bias is (131), as we expect.

The variances are equal to the negative inverses of the expected Hessians. These, combined with bias (131), readily leads to the MSE. Of course, (145) and (148) are for $\hat{\alpha}_c$ and hence the delta method needs to be applied to obtain the variances for $\hat{\pi}_c$. Note that the variance

for $\hat{\pi}$ was already derived in (130), but applying the delta method to (141) gives the exact same result. The additional expressions are

$$\text{var}(\hat{\pi}_c|N = n) = \frac{\pi(1 - \pi)}{n - 1}, \quad (150)$$

$$\text{var}(\hat{\pi}_c|N = 2n) = \frac{\pi(1 - \pi)}{2n - 1}, \quad (151)$$

with the expected conditional Hessians the inverses of these quantities:

$$\text{var}(\hat{\pi}_c) = \text{var}(\hat{\pi}_c|N = n) \cdot \pi + \text{var}(\hat{\pi}_c|N = 2n) \cdot (1 - \pi) = \frac{(n\pi + n - 1)}{(n - 1)(2n - 1)}\pi(1 - \pi). \quad (152)$$

Note that the derivation of overall variance (152) involves the expectation of the conditional variances only, while the variance of the conditional expectations is zero, because both conditional estimators are unbiased. Finally,

$$\begin{aligned} \text{MSE}(\hat{\pi}) &= \text{var}(\hat{\pi}) + \text{Bias}(\hat{\pi}) = \frac{\pi(1 - \pi)}{n(2 - \pi)} + \frac{\pi^2(1 - \pi)^2}{(2n)^2} = \frac{1}{n} \cdot \frac{\pi(1 - \pi)}{2 - \pi} + \mathcal{O}(n^{-2}), \quad (153) \\ \text{MSE}(\hat{\pi}_c) &= \text{var}(\hat{\pi}_c) = \frac{(n\pi + n - 1)\pi(1 - \pi)}{(n - 1)(2n - 1)} = \frac{1}{n} \cdot \frac{\pi(1 - \pi)(\pi + 1)}{2} + \mathcal{O}(n^{-2}). \quad (154) \end{aligned}$$

Calculating the difference between (154) and (153), we find

$$\text{MSE}(\hat{\pi}_c) - \text{MSE}(\hat{\pi}) = \frac{1}{n} \frac{\pi^2(1 - \pi)^2}{2(2 - \pi)} + \mathcal{O}(n^{-2}).$$

Hence, like in the normal case, the joint estimator is more efficient than the marginal one. Of course, the MSE increase when moving from the joint to the conditional estimator is modest, with $\text{MSE}(\hat{\pi}) \leq \text{MSE}(\hat{\pi}_c) \leq 1.125\text{MSE}(\hat{\pi})$, the maximum discrepancy reached for $\pi = 0.5$, and equality for $\pi = 0$ or $\pi = 1$.

6 Concluding Remarks

We have considered the consequences for statistical inference of a random sample size. Our setting is that of univariate random variables from the exponential family that are subject to a stopping rule such that the sample size is either $N = n$ or $N = 2n$, with n

specified by design. The stopping rule is stochastic and is allowed to depend on the sample sum K over the first n observations. The rule is generic in the sense that its limiting cases are a deterministic stopping rule, such as in a sequential trial, and a completely random sample size, independent of the data. This setting extends those of both Liu *et al* (2006) and Molenberghs *et al* (2012); the former restrict attention to a deterministic stopping rule, although they do so for an arbitrary number of interim looks. The latter confined attention to normally distributed outcomes only.

We have focused on three important inferential aspects. First, we have shown that the sufficient statistic (K, N) is incomplete. Second, we have examined the consequences of this for the sample average, as well as for linear generalizations thereof. We have shown that there is small-sample bias, except for the CRSS case. Even then, there is no optimal estimator, except for the exponential distribution, for which the optimum differs from the ordinary sample average. Third, we have studied maximum likelihood estimation in both a joint as well as a conditional framework. The joint likelihood is for the exponential-family parameter and the stopping rule simultaneously. The conditional likelihood starts from the conditional distribution of the outcomes, given the sample size. Also here, counterintuitive results are derived. The joint likelihood produces the sample average as maximum likelihood estimator, which is biased in finite samples but is asymptotically unbiased, provided a regularity condition on the stopping rule applies. The conditional likelihood estimator is unbiased, even in small samples. This notwithstanding, the sample average has smaller MSE than the conditional estimator in many important cases, such as the normal and binary examples considered, as well as when the variance of the outcomes is sufficiently small. Under regularity conditions, both estimators are asymptotically equivalent, with the difference between both being $\mathcal{O}(n^{-1})$. The regularity condition is not very restrictive; it essentially comes down to requiring that $F'(k = n\mu)$ approaches zero where F is the stopping rule. For broad classes of parametric functions, this condition

is satisfied. We have shown that the corresponding conditional expectations are unbiased. Hence, when the regularity conditions are satisfied, the sample average remains an attractive and sensible choice for sequential trials. Thus, while some familiar inferential properties no longer hold, estimation after sequential trials is more straightforward than commonly considered and there is little need for complicated, modified estimators, given that the ordinary sample average is acceptable for wide classes of stopping rules, whether stochastic or deterministic.

Molenberghs *et al* (2012) considered several ramifications of their developments. They commented on the situation of an arbitrary number of looks in a sequential trial, and considered in detail the CRSS case for more than two possible sample sizes. All of this was done for normally distributed outcomes. They also commented on the connection between their derivations and longitudinal outcomes subject to dropout of an MAR type, where dropout depends on observed but not further on unobserved outcomes. While similar, there are subtle differences because now the randomness in the sample size pertains to the number of measurements per subject, rather than to the number of subjects. The difference lies in the fact that measurements within a subject are not independent. Our results extend to these settings as well for the exponential family. Furthermore, connections can be made with a variety of other settings with random sample sizes, such as clustered data with informative cluster sizes, time-to-event data subject to censoring, jointly observed longitudinal and time-to-event data, and random observation times. These settings are currently scrutinized further, and will be reported in a separate manuscript.

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.

References

- Armitage, P. (1975) *Sequential Medical Trials*. Oxford: Blackwell.
- Barndorff-Nielsen, O. and Cox, D.R. (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review*, **52**, 309–326.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, **15**, 377–380.
- Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall.
- Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, **77**, 875–892.
- Grambsch, P. (1983). Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Annals of Statistics*, **11**, 68–77.
- Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.
- Hughes, M.D. and Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine*, **7**, 1231–1242.
- Jennison, C. and Turnbull, B.W. (2000). *Group Sequential Methods With Applications to Clinical Trials*. London: Chapman & Hall/CRC.
- Kenward, M.G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **13**, 236–247.
- Lehmann, E.L. and Stein, C. (1950) Completeness in the sequential case. *Annals of Mathematical Statistics*, **21**, 376–385.
- Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.
- Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.

- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- Liu, A. and Hall, W.J. (1999). Unbiased estimation following a group sequential test. *Biometrika*, **86**, 71–78.
- Liu, A., Hall, W.J., Yu, K.F., and Wu, C. (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family. *Statistica Sinica*, **16**, 165–81.
- Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2012). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, **00**, 000–000.
- Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, **75**, 723–729.
- Rotnitzky, A. (2009). Inverse probability weighted methods. In: *Longitudinal Data Analysis* (G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs eds.), 453–476. CRC/Chapman & Hall, Boca Raton.
- Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, **66**, 688–701.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.
- Siegmund, D. (1978). Estimation following sequential tests. *Biometrika*, **64**, 191–199.
- Tsiatis, A.A., Rosner, G.L., and Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*, **40**, 797–803.
- Todd, S., Whitehead, J., and Facey, K.M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika*, **83**, 453–461.
- Vansteelandt, S., Carpenter, J.R., and Kenward, M.G. (2010) Analysis of incomplete data using inverse probability weighting and doubly robust estimators *Methodology*, **6**, 37–48.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, **16**, 117–186.
- Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials (2nd ed.)*. New York: John Wiley & Sons.
- Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine*, **18**, 2271–2286.

A Stopping Probability for Normally Distributed Outcomes

Molenberghs *et al* (2012) derived (22) from first principles. It is also possible to derive this expression from (9), by plugging in the standard normal density for $\tilde{f}_1(z)$ with $A(k) = \alpha + \beta k/n$, combined with $f_{n,\theta=\mu} = \varphi_{\mu,n}(k)$.

With these choices, we find:

$$f(N = n) = \exp\left(-\frac{n}{2}\mu^2\right) \cdot I$$

with

$$\begin{aligned} I &= \int_{k=-\infty}^{k=+\infty} \frac{e^{\mu k - \frac{1}{2}\frac{k^2}{n}}}{\sqrt{n}\sqrt{2\pi}} \int_{z=-\infty}^{z=\alpha+\beta k/n} \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} dz dk \\ &= \frac{1}{\sqrt{n}(\sqrt{2\pi})^2} \int_{k=-\infty}^{k=+\infty} \int_{t=-\infty}^{t=\alpha} \exp\left[\mu k - \frac{1}{2}k^2n - \frac{1}{2}(t + \beta k/n)^2\right] dt dk \\ &= \frac{1}{\sqrt{n}(\sqrt{2\pi})^2} \int_{k=-\infty}^{k=+\infty} \int_{t=-\infty}^{t=\alpha} \exp\left[-\frac{1}{2}\frac{(k-\ell)^2}{q^2} - \frac{1}{2}\frac{(t-L)^2}{Q^2} - \frac{1}{2}M\right] dt dk, \quad (155) \end{aligned}$$

with

$$\begin{aligned} \ell &= \frac{n\mu - \beta t}{\beta^2 + n} \cdot n, \\ q &= \frac{n}{\sqrt{\beta^2 + n}}, \\ L &= -\mu\beta, \\ Q &= \frac{\sqrt{\beta^2 + n}}{\sqrt{n}}, \\ M &= -\mu^2 n. \end{aligned}$$

This implies that

$$f(N = n) = \exp\left(-\frac{n}{2}\mu^2\right) \cdot \frac{1}{\sqrt{n}} \cdot q \cdot Q \cdot \Phi\left(\frac{\alpha - L}{Q}\right) \exp\left(+\frac{n}{2}\mu^2\right) = \Phi\left(\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}}\right),$$

which equals (22).

B Joint Probability for Binary Outcome with Beta-driven Stopping Rule

Considering (44), expression (48) is valid if we can show that

$$\frac{2n-k}{2n} \binom{2n}{k} = \binom{2n}{k} - H(k), \quad (156)$$

with

$$H(k) = \sum_{z=0 \vee (k-n)}^{k \wedge n} \binom{n}{z} \binom{n}{k-z} \frac{z}{n}. \quad (157)$$

In other words, we need to show

$$\frac{k}{2n} \binom{2n}{k} = H(k),$$

i.e.,

$$\binom{2n-1}{k-1} = \sum_{z=1 \vee (k-n)}^{k \wedge n} \binom{n-1}{z-1} \binom{n}{k-z} = \sum_{z'=0 \vee (k-1-n)}^{k-1 \wedge n-1} \binom{n-1}{z'} \binom{n}{k-1-z'},$$

which is obviously correct.

C Conditional Expectations for the Conditional Likelihood in the Binary Case

As stated in Section 5.3, it is easy to show from first principles and instructive that the conditional expectations of the conditional scores (134) and (137).

An explicit expression for $A'_n(\pi)$ in (134) is

$$\begin{aligned} A'_n(\pi) &= \sum_{k=0}^n \binom{n}{k} k \pi^{k-1} (1-\pi)^{n-k} F(k) - \sum_{k=0}^n \binom{n}{k} n \pi^k (1-\pi)^{n-k-1} F(k) \\ &\quad + \sum_{k=0}^n \binom{n}{k} k \pi^k (1-\pi)^{n-k-1} F(k). \end{aligned}$$

The expectation becomes

$$\begin{aligned}
E[S_n(\pi)|N = n] &= E\left[\frac{k}{\pi(1-\pi)} - \frac{n\pi}{\pi(1-\pi)} - \frac{A'_n(\pi)}{A_n(\pi)}\right] \\
&= \frac{\sum_{k=0}^n \binom{n}{k} k\pi^k(1-\pi)^{n-k} F(k)}{A_n(\pi)\pi(1-\pi)} - \frac{n\pi}{\pi(1-\pi)} \\
&\quad - \frac{\sum_{k=0}^n \binom{n}{k} k\pi^k(1-\pi)^{n-k} F(k)}{A_n(\pi)\pi(1-\pi)} \cdot (1-\pi) \\
&\quad + \frac{n \sum_{k=0}^n \binom{n}{k} \pi^k(1-\pi)^{n-k} F(k)}{A_n(\pi)\pi(1-\pi)} \cdot \pi \\
&\quad - \frac{\sum_{k=0}^n \binom{n}{k} k\pi^k(1-\pi)^{n-k} F(k)}{A_n(\pi)\pi(1-\pi)} \cdot \pi \\
&= \frac{P(\pi)}{A_n(\pi)\pi(1-\pi)} - \frac{n\pi}{\pi(1-\pi)} - \frac{P(\pi)}{A_n(\pi)\pi(1-\pi)} \cdot (1-\pi) \\
&\quad + \frac{n}{\pi(1-\pi)} \cdot \pi - \frac{P(\pi)}{A_n(\pi)\pi(1-\pi)} \cdot \pi \\
&= 0,
\end{aligned}$$

where

$$P(\pi) = \sum_{k=0}^n \binom{n}{k} k\pi^k(1-\pi)^{n-k} F(k)$$

Because (137) is slightly more involved than (134), we proceed differently. Writing

$$S_{2n}(\pi) = \frac{1}{\pi(1-\pi)} \left\{ K - \frac{2n\pi}{1-A(K,\pi)} + \frac{\sum_{\ell=0}^{2n} \ell\pi^\ell(1-\pi)^{2n-\ell} \sum_z \binom{n}{z} \binom{n}{\ell-z} F(z)}{1-A(\pi)} \right\},$$

with

$$A(\pi) = \sum_{\ell=0}^{2n} \pi^\ell(1-\pi)^{2n-\ell} \sum_z \binom{n}{z} \binom{n}{\ell-z} F(z),$$

and using the result that

$$2n\pi = \sum_{k=0}^{2n} k\pi^k(1-\pi)^{2n-k},$$

we can write the score as:

$$S_{2n}(\pi) = \frac{1}{\pi(1-\pi)} \left\{ K - \frac{\sum_{\ell=0}^{2n} K p(\ell, 2n)}{1 - A(K, \pi)} \right\} = \frac{1}{\pi(1-\pi)} \left\{ K - \frac{\sum_{\ell=0}^{2n} K p(\ell, 2n)}{p(N = 2n)} \right\}$$

which immediately implies that $E[S_{2n}(\pi)|N = 2n] = 0$.