

A Tutorial on the Practical Use and Implication of Complete Sufficient Statistics

Peer-reviewed author version

HERMANS, Lisa; MOLENBERGHS, Geert; AERTS, Marc; Kenward, Michael G. & VERBEKE, Geert (2018) A Tutorial on the Practical Use and Implication of Complete Sufficient Statistics. In: INTERNATIONAL STATISTICAL REVIEW, 86(3), p. 403-414.

DOI: 10.1111/insr.12261

Handle: <http://hdl.handle.net/1942/27561>

A Tutorial on the Practical Use and Implication of Complete Sufficient Statistics

Lisa Hermans¹, Geert Molenberghs^{1,2}, Marc Aerts¹,
Michael G. Kenward³ and Geert Verbeke^{2,1}

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

² *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

³ *Luton, United Kingdom*

e-mail: lisa.hermans@uhasselt.be

Summary

Completeness means that any measurable function of a sufficient statistic that has zero expectation for every value of the parameter indexing the parametric model class, is the zero function almost everywhere. The property is satisfied in many simple situations in view of parameters of direct scientific interest, such as in regression models fitted to data from a random sample with fixed size. A random sample is not always of a fixed, a priori determined size. Examples include sequential sampling and stopping rules, missing data, and clusters with random size. Often there then is no complete sufficient statistic. A simple characterization of incompleteness is given for the exponential family in terms of the mapping between the sufficient statistic and the parameter, based upon the implicit function theorem. Essentially this is a comparison of the dimension of the sufficient statistic to the length of the parameter vector. This results in an easy verifiable criterion for incompleteness, clear and simple to use, even for complex settings as is shown for missing data and clusters of random size.

This tutorial exemplifies the (in)completeness property of a sufficient statistic, thereby illustrating our proposed characterization. The examples are organized from more classical, simple examples to gradually more advanced settings.

Key words: Ancillarity, censoring, incomplete data, joint modeling, random cluster size, sequential trial.

1 Introduction

The simplest statistical designs involve the collection of a univariate or multivariate outcome, often with an accompanying vector of covariates, for a number of independent study units.

This number, the sample size, is classically fixed *a priori*. However, many designs frequently deviate from this in important ways. One consequence is that *complete* sufficient statistics may no longer exist. Completeness implies that any function of a sufficient statistic that has zero expectation for every value of the parameter indexing the parametric model class, is the zero function almost everywhere. The relevance of complete sufficient statistics has been established through two theorems, Lehman-Scheffé theorem and Basu's theorem. Completeness, combined with regularity conditions, provides a basis for estimators with desirable properties, such as unbiasedness and optimality, discussed further. Note, this paper confines to the (in)completeness of sufficient statistics, excluding statistics that are not sufficient.

In sequential designs (Wald, 1945) one incorporates a data-driven rule to potentially stop the trial before reaching the maximal sample size. Such methods are well established in clinical trials (Armitage, 1975). While the statistical aspects of sequential methods have been carefully studied (Lehman and Stein, 1950), the lack of completeness has led to disagreement and confusion, regarding appropriate (point and interval) estimation following such trials, leading to many *ad hoc* proposals. Liu and Hall (1999) and Liu *et al.* (2006), building upon Emerson and Fleming (1990), explored this aspect. Molenberghs *et al.* (2014) and Milanzi *et al.* (2014, 2015) studied the issue in a wider framework, encompassing stochastic stopping rules and completely random sample sizes. They demonstrated that, somewhat contrary to intuition and in spite of incompleteness, the ordinary sample average remains a viable estimator (because of consistency, asymptotic normality, and high efficiency), even though it no longer has all properties that it enjoys in the conventional, fixed sample size setting. We elaborate on this in Section 2.1. Another setting without complete sufficient statistics is that of clusters of unequal size. Such designs include longitudinal, multilevel, spatial, and multi-stage survey designs. A counterexample is a longitudinal study where each one of the subjects is measured exactly the same number of times, at an *a priori* fixed set of measurement occasions. Then, N , the number of subjects, and n , the amount of measurements, are design constants. However, such “clean” designs are the exception rather than the norm. A variety of *ad hoc* methods has been proposed for the random cluster size setting. Other settings without complete sufficient statistics are missing data, censored time-to-event data, random visit times, and joint modeling

of longitudinal and time-to-event data.

Although the definition of a complete sufficient statistic is clear, its constructive verification in a given situation often involves tedious algebra. This is especially true in sequential trials, except for the simplest situation of two possible sample sizes only; such calculations are, quite literally, convoluted. Likewise, when completeness does not hold, the construction of counterexamples may or may not be straightforward. Nevertheless, a clear, simple, and easily verifiable criterion for completeness, of a constructive rather than an existential nature, would be welcome. For example, in a normal univariate sample with fixed sample size, a minimal sufficient statistic for the population mean is the sample sum, in contrast to the random sample case for which it is the sample sum *and* the realized (random) sample size. The parameter remains one-dimensional, but the minimal sufficient statistic is two-dimensional, and eventually leading to incompleteness. A general criterion can be formulated that starts from, but moves beyond, the length of a vector.

To ensure completeness of the minimal sufficient statistics Lehmann (1981, pp. 142–143), Brown (1986, pp. 42–44) and Boos and Stefanski (2013, pp. 103–104) formulated theorems, based on appropriate restrictions placed on the canonical form of the exponential family. Brown (1986) proves incompleteness using complex analytic properties and refers to the unique determination of a standard family by its Laplace transform. In this paper, however, the latter is more explicitly used and a result, both general and easy to use, follows. Boos and Stefanski (2013) and Lehmann (1981) base their theorems on the fact that the family is minimal and the parameter space contains a rectangle, thereby requiring that the family is of full rank. The characterization of incompleteness given in this paper is also related to a property of curved exponential models (Van Garderen, 1997; Keener, 2010). These have the property that the dimension of the minimal sufficient statistic is larger than the number of parameters in the model. Van Garderen (1997) establishes a theorem that allows a straightforward comparison between the dimension of the minimal sufficient statistic and the number of parameters to determine when a model is a curved exponential model. Keener (2010) points out that curved exponential models arise naturally with data from sequential experiments and in applications to contingency table analysis.

In Section 2, two commonly encountered settings are presented, where minimal sufficient statistics are incomplete. Known results leading up to the characterization of complete sufficient statistics are briefly reviewed in Section 3. The key result, our characterization, is presented in Section 4. To highlight the ease of use of the criterion, it is applied and shown to work for two more complex data settings, i.e., clusters of random size and missing data. Section 5 illustrates and further clarifies our findings for clustered data. Section 6 considers partially unobserved contingency tables, extends these results to other missing-data settings and shows why seemingly unrelated settings, all have led to incomplete sufficient statistics.

2 Motivating Settings

2.1 Sequential Trials

Group sequential trials are in common use and have been well studied (e.g., Wald, 1945; Armitage, 1975; Whitehead, 1997; Jennison and Turnbull, 2000). The corresponding design and hypothesis testing machinery is well developed. Nevertheless, issues still surround estimation following a sequential trial (Siegmund, 1978; Hughes and Pocock, 1988; Todd, Whitehead, and Facey, 1996; Whitehead, 1999). Several authors have reported that standard estimators such as the sample average are biased. In response to this, various proposals have been made to remove or alleviate this bias and its consequences (Tsiatis, Rosner, and Mehta, 1984; Rosner and Tsiatis, 1988; Emerson and Fleming, 1990). An early suggestion was to use an estimator (Blackwell, 1947) that conditions on the stopping event.

The origin of the problem was understood at an early stage of the development. Lehman and Stein (1950) showed that it originates from *incompleteness* of the sufficient statistics, generally implying the non-existence of a minimum variance unbiased linear estimator. Liu and Hall (1999) and Liu *et al.* (2006) explored this incompleteness in group sequential trials, and Molenberghs *et al.* (2014) and Milanzi *et al.* (2014, 2015) embedded the problem in the broader class of random sample size, which also includes, missing data, completely random sample sizes, censored time-to-event data, and random cluster sizes. Their main findings were:

(1) the sample average, although asymptotically unbiased has finite sample bias; (2) apart from the exponential distribution setting, there is no finite-sample optimal linear estimator, although the sample average is asymptotically optimal (i.e., uniform minimum variance unbiased); (3) the validity (i.e., consistency and asymptotic normality) of the sample average also follows from standard ignorable likelihood theory (Little and Rubin, 2002); we will return to ignorability in Section 6; (4) there exists a maximum likelihood estimator that conditions on the realized sample size, which is finite sample unbiased, but has slightly larger variance and mean square error.

2.2 Clusters of Unequal Size

While given less attention, there is an extensive literature on what is often called ‘informative cluster size,’ taken to mean that the cluster size contains some information about the parameters of scientific interest, which should be contrasted with the use of the term ‘informative’ in the missing-data and event-time literatures. Even when the cluster size contains no information about the scientific parameters, there are issues resulting from lack of a complete sufficient statistic.

One family of approaches is based on restricted moment estimators obtained through the use of generalized estimating equations (Liang and Zeger, 1986; Liang, Zeger, and Qaqish, 1991). Pseudo-likelihood, or composite likelihood, estimators have also been proposed (Lindsay, 1988; Arnold and Strauss, 1991; le Cessie and van Houwelingen, 1994; Geys, Molenberghs, and Lipsitz, 1998; Aerts *et al.*, 2002). In these, the full likelihood is simplified and replaced by a more manageable function (Geys, Molenberghs, and Lipsitz, 1998). Various authors have studied weighted and unweighted approaches, whereas (non-)informative cluster sizes (Williamson, Datta, and Satten, 2003; Benhin, Rao, and Scott, 2005; Hofman, Sen, and Weinberg, 2001; Cong, Yin, and Shen, 2007; Chiang and Lee, 2008; Wang, Kong, and Datta, 2011).

3 (In)complete Sufficient Statistics and Some Known Results

The property of central interest is that of *completeness* (Casella and Berger, 2001, pp. 285–286). A statistic $k(Y)$, a measurable function of a random variable Y and with Y belonging to a family P_θ , is complete if, for every measurable function $g(\cdot)$, independent of θ , the property $E[g\{k(Y)\}] = 0$ for all θ , implies that $P_\theta[g\{k(Y)\} = 0] = 1$ for all θ . The relevance of completeness rests on two follow-up theorems. First, the Lehman-Scheffé theorem (Casella and Berger, 2001) states that, if a statistic is unbiased, complete, and sufficient for a parameter θ , then it corresponds to the best mean-unbiased estimator for θ . Second, the connection with ancillarity follows from Basu's theorem (Basu, 1955): a statistic that is both bounded complete and sufficient is independent of any ancillary statistic (Casella and Berger, 2001, p. 287). The theorems are implications rather than equivalences. For example, in the sequential trial context there exist estimators with very good properties, despite lack of completeness (Molenberghs *et al.*, 2014).

Liu and Hall (1999) established incompleteness of the sufficient statistic for a clinical trial with a stopping rule, for the case of normally distributed outcomes. Liu *et al.* (2006) generalized this result to the exponential family. Molenberghs *et al.* (2014) and Milanzi *et al.* (2014) broadened it further to a stochastic stopping rule, encompassing the important case of a completely random sample size. In the latter case, even though sample size and data are unrelated, completeness no longer holds.

Tables 1 and 2 contain a number of illustrative examples where the sufficient statistics are found to be (in)complete. In Table 1, continuous and categorical outcomes are considered. Positive outcomes (continuous times and counts) are the subject of Table 2. Some of these models are based upon Chakraborty (2015). Precise formulations and details can be found in Supplementary Materials A. Examples 1 and 2, a univariate sample with either known or unknown variance, have complete sufficient statistics. Example 3, a univariate normal sample with coupled mean and variance, does not; here, unlike in the previous examples, the sufficient statistic is of higher dimension than the parameter. When the mean-variance coupling parameter τ^2 is unknown (Example 3a), the sufficient statistic and the parameter are

again of the same dimension and completeness holds, unlike when τ^2 is known (Example 3b). If the statistic is restricted to either the sample sum or the sum of squared sample units, then it is no longer sufficient. This last situation occurs also in Example 4, a sequential trial, where the sufficient statistic consist not only of the data collected, but also of the sample size realized, i.e., a one-dimensional parameter needs a two-dimensional sufficient statistic. These developments emphasize that the establishment of either completeness or its converse requires tedious, situation-specific calculations when using the definition. It is therefore convenient to derive a simple criterion based on the dimensions of the parameter vector and the sufficient statistic, to be established next.

4 A Characterization of Incompleteness

We turn to a general characterization of incompleteness, in the exponential family with a vector-valued parameter and minimal sufficient statistic. Group the outcomes Y_i into a vector \mathbf{Y} , with vector-valued parameter $\boldsymbol{\theta}$ and write the exponential family in the form

$$f(\mathbf{y}|\boldsymbol{\theta}) = \tilde{h}(\mathbf{y}) \exp \{ \boldsymbol{\eta}(\boldsymbol{\theta})' \mathbf{k}(\mathbf{y}) - A(\boldsymbol{\theta}) \}, \quad (1)$$

where the sufficient statistic $\mathbf{K} \equiv \mathbf{K}(\mathbf{Y})$. Consider first the situation where the function $\boldsymbol{\eta}$ is everywhere of full rank. Examples 1 and 2 fall into this category. Because $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are in 1-to-1 relationship, we can use $\boldsymbol{\theta}$, without loss of generality. The score equation corresponding to (1) is $S(\boldsymbol{\theta}) = \partial \boldsymbol{\eta} / \partial \boldsymbol{\theta} \cdot \mathbf{K} - \partial A / \partial \boldsymbol{\theta} = 0$. If the transformation, $\boldsymbol{\eta}(\boldsymbol{\theta})$, is of full rank, then it follows that

$$\mathbf{K} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}} \right)^{-1} \frac{\partial A}{\partial \boldsymbol{\theta}}. \quad (2)$$

Taking expectations, the right hand side of Equation (2) equals $E(\mathbf{K})$. In the above situation, the sufficient statistic is complete. To see this, assume that there is a function $g(\mathbf{k})$ with expectation zero for all values of $\boldsymbol{\theta}$. It then satisfies

$$\int g(\mathbf{k}) h(\mathbf{k}) \exp \{ \boldsymbol{\theta}' \mathbf{k} - A(\boldsymbol{\theta}) \} d\mathbf{k} = 0, \quad (3)$$

Table 1: Examples with complete and incomplete sufficient statistics (continuous and categorical outcomes)

<i>Ex.</i>	Setting	Parameter(s)	Sufficient statistic(s)
Settings with complete sufficient statistics			
1	$Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$ with μ unknown and σ^2 known	μ	K_1
2	$Y_i \sim N(\mu, \sigma^2), i = 1, \dots, n$ with μ and σ^2 unknown	(μ, σ^2)	(K_1, K_2)
3a	$Y_i \sim N(\mu, \tau^2 \mu^2), i = 1, \dots, n$ with μ and τ^2 unknown	(μ, τ^2)	(K_1, K_2)
6	$Y_i \sim N(\mu, \mu), i = 1, \dots, n$	μ	K_2
7a	$Y_i \sim N(\mu, \mu^{2\lambda}), i = 1, \dots, n$ and $\lambda = 0$ or $1/2$	μ	K_1 or K_2
8	$M_1 \times M_2$ contingency table with $\phi(k_1 k_2)$ and $\pi(k_2)$	$\varphi(k_1 k_2), \pi(k_2)$	
15	Fully observed 2×1 contingency table	p	Z_{21}
Settings with incomplete sufficient statistics			
3b	$Y_i \sim N(\mu, \tau^2 \mu^2), i = 1, \dots, n$ with μ unknown and τ^2 known	μ	(K_1, K_2)
4	Sequential trial with stochastic stopping rule	μ	(K_3, N)
5	Bivariate parameter, one of which known (cf. Ex. 2)	μ	(K_1, K_2)
7b	$Y_i \sim N(\mu, \mu^{2\lambda}), i = 1, \dots, n$ and $\lambda \neq 0$ and $\neq 1/2$	μ	K_1, K_2
9	$Y_i \sim N(\mu, 1)$, sample size N , $1 \leq N \leq n$ with π_N	μ	(K_3, N)
10	$\mathbf{Y} \sim N(\mu \mathbf{1}_N, \sigma^2 I_N + \tau^2 J_N)$	(μ, σ^2)	(K_3, K_4, K_5, N)
11	Vector-valued data and parameter, with completely random sample size	$\pi(N k)$	(K_3, N)
12	N clusters of completely random size		$[\mathbf{K} = \mathbf{K} \{(\mathbf{Y}_i)\};$ $\mathbf{N} = \mathbf{N} \{(N_i)\}]$
13	$\mathbf{Y} \sim N(\mu \mathbf{1}_N, \sigma^2 I_N + \tau^2 J_N), i = 1, \dots, N$	(μ, σ^2, τ^2)	$(S_{1\ell}, S_2, S_{3\ell}, S_{4\ell})$
14	General clustered-data setting with random cluster sizes	θ	
16	Partially missing 2×1 contingency table	p	(Z_{21}, Z_1)
17	Partially missing 2×1 contingency table	p_{jk}	(Z_{2jk}, Z_{1j})

$$K_1 = \sum_{i=1}^n Y_i; K_2 = \sum_{i=1}^n Y_i^2; K_3 = \sum_{i=1}^N Y_i; K_4 = \mathbf{Y}'\mathbf{Y}; K_5 = \mathbf{Y}'J_N\mathbf{Y}.$$

$$S_{1\ell} = \sum_{i=1}^{c_\ell} \sum_{j=1}^{n_\ell} Y_{ij}^{(\ell)}; S_2 = \sum_{\ell=1}^L \sum_{i=1}^{c_\ell} \sum_{j=1}^{n_\ell} \left(Y_{ij}^{(\ell)}\right)^2; S_{3\ell} = \sum_{i=1}^{c_\ell} \left(\sum_{j=1}^{n_\ell} Y_{ij}^{(\ell)}\right)^2; S_{4\ell} = c_\ell.$$

Table 2: Examples with complete and incomplete sufficient statistics (outcomes on $[0, +\infty[$)

<i>Ex.</i>	Setting	Parameter(s)	Sufficient statistic(s)
Settings with complete sufficient statistics			
18	$Y_i \sim \text{Poisson}(\lambda)$	λ	K_1
19	$Y_i \sim \text{Exponential}(\lambda)$	λ	K_1
20	$Y_i \sim \text{Integrated Exponential}(\lambda)$	λ	K_1
Settings with incomplete sufficient statistics			
21	$Y_i \sim \text{Integrated Weibull}(\lambda, \rho)$	(λ, ρ)	Y_1, \dots, Y_n
$K_1 = \sum_{i=1}^n Y_i.$			

with obvious notation, similar to Equation (1) but \tilde{h} expressed as function of \mathbf{k} rather than \mathbf{y} . Applying Fubini's theorem (Rudin, 1974), we can write Equation (3) as

$$0 = \int dk_p h(k_p) e^{\theta_p k_p} \int dk_{p-1} h(k_{p-1} | k_p) e^{\theta_{p-1} k_{p-1}} \dots$$

$$\int dk_1 g(k_1, \dots, k_p) h(k_1 | k_2, \dots, k_p) e^{\theta_1 k_1}.$$

This leads to a telescopic series of Laplace transforms:

$$F_{\theta_1}(k_2, \dots, k_p) = \mathcal{L}_{\theta_1} \{g(k_1, \dots, k_p) h(k_1 | k_2, \dots, k_p)\}, \quad (4)$$

$$F_{\theta_2}(k_3, \dots, k_p) = \mathcal{L}_{\theta_2} \{F_{\theta_1}(k_2, \dots, k_p) h(k_2 | k_3, \dots, k_p)\}, \quad (5)$$

$$\vdots$$

$$F_{\theta_p} = \mathcal{L}_{\theta_p} \{F_{\theta_{p-1}}(k_p) h(k_p)\} = 0, \quad (6)$$

with obvious notation. Moving step-by-step from Equations (6) to (4), the sequence of F_{θ_j} is zero a.e. for j running down from $p-1$ to 1 and then eventually $g(k_1, \dots, k_p) = 0$ a.e., establishing completeness. Looking to the bivariate Examples 2 and 5, the same is seen for Example 2, but a different result occurs for Example 5 where one of the two parameters is known. Then, the sufficient statistic is incomplete.

Proposition 1. (*Characterization of a complete sufficient statistic.*) *Provided the parameter space is rectangular, a sufficient statistic \mathbf{k} is complete for a parameter $\boldsymbol{\theta}$ in a*

exponential family model if and only if $\boldsymbol{\theta}$ cannot be transformed to a parameterization $\boldsymbol{\eta}$ with a proper subset $\boldsymbol{\eta}_1$ such that $\boldsymbol{\eta} = (\boldsymbol{\eta}'_1, \boldsymbol{\eta}_2(\boldsymbol{\eta}_1)')'$.

Proof. The proof is based upon a more general version of Example 5.

Let $\boldsymbol{\eta}(\boldsymbol{\theta})$ be a function to match the minimal sufficient statistic \mathbf{K} that is not of full rank. This can be decomposed, using the implicit function theorem, assuming the functions involved are continuously differentiable, and then mapped as follows:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2(\boldsymbol{\eta}_1) \end{pmatrix} \longleftrightarrow \mathbf{K} = \begin{pmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{pmatrix}. \quad (7)$$

Incompleteness of the sufficient statistic would follow if a function $g(\mathbf{k}_1, \mathbf{k}_2)$ could be found satisfying:

$$0 = \int d\mathbf{k}_2 h_2(\mathbf{k}_2) e^{\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)' \mathbf{k}_2} \int d\mathbf{k}_1 h_1(\mathbf{k}_1 | \mathbf{k}_2) g(\mathbf{k}_1, \mathbf{k}_2) e^{\boldsymbol{\eta}_1' \mathbf{k}_1}, \quad (8)$$

$$0 = \int d\mathbf{k}_2 h_2(\mathbf{k}_2) e^{\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)' \mathbf{k}_2} \mathcal{L}_{\boldsymbol{\eta}_1} \{h_1(\mathbf{k}_1 | \mathbf{k}_2) g(\mathbf{k}_1, \mathbf{k}_2)\}, \quad (9)$$

$$0 = \int h_2(\mathbf{k}_2) F(\mathbf{k}_2, \boldsymbol{\eta}_1) e^{\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)' \mathbf{k}_2} d\mathbf{k}_2. \quad (10)$$

Now, (10) is not a Laplace transform. Therefore, we can choose a function $F(\mathbf{k}_2, \boldsymbol{\eta}_1)$ that satisfies the equation and then use the inverse Laplace transform to derive $g(\mathbf{k}_1, \mathbf{k}_2)$. To see that such a function can easily be found, choose:

$$F(\mathbf{k}_2, \boldsymbol{\eta}_1) = e^{-\boldsymbol{\eta}_2(\boldsymbol{\eta}_1)' \mathbf{k}_2} \tilde{F}(\mathbf{k}_2).$$

With this choice, condition (10) simplifies to:

$$0 = \int h_2(\mathbf{k}_2) \tilde{F}(\mathbf{k}_2) d\mathbf{k}_2.$$

In other words, we need a function $\tilde{F}(\mathbf{k}_2) \perp h_2(\mathbf{k}_2)$. □

Notice the similarity between the characterization and earlier work of Lehmann (1981), Brown (1986) and Boos and Stefanski (2013). However, our characterization leads to more general result and easy to use criterion. Also Van Garderen (1997) and Keener (2010) have already pointed out this relationship between the dimension of the sufficient statistic and the number of parameters for curved exponential models. Evidently, their focus is different from ours. With this characterization all examples of Tables 1 and 2 can be verified solely by counting the dimensions of the parameter vectors and sufficient statistic. The proposition explains why Examples 1 and 2 have complete sufficient statistics. This is trivial in Example 1 because the parameter and sufficient statistic are scalar. In Example 2, the parameter $\theta = (\mu, \sigma^2)'$ consists of two functionally independent components. Example 3 has a bivariate sufficient statistic, like Example 2, and a bivariate parameter $\theta = (\mu, \tau^2 \mu^2)'$. Write $\eta_1 = \mu$ and $\eta_2(\eta_1) = \tau^2 \eta_1^2$, which explains why this is an incomplete case when τ^2 is known. For Example 4, consider two sample sizes n and $2n$. The minimal sufficient statistic is (K_3, N) , and both are governed by a distribution with sole parameter μ , trivially establishing incompleteness. This result relates to Shao (1999, p. 110). In their Proposition 2.1, they consider the exponential family case, of full rank, for a sufficient statistic that is complete and sufficient. Their proof is in terms of the positive and negative parts of the normalizing function, rather than Laplace transforms.

Corollary 1. *(Non-linearity of the function $\eta_2(\eta_1)$.) For complete minimal sufficient statistics, the function $\eta_2(\eta_1)$ cannot be linear.*

Proof. To see this, assume there is such a linear function. The correspondence becomes:

$$\eta(\theta) = \begin{pmatrix} \eta_1 \\ L\eta_1 \end{pmatrix} \longleftrightarrow \mathbf{K} = \begin{pmatrix} \mathbf{K}_1 \\ \mathbf{K}_2 \end{pmatrix}.$$

The inner product of these vectors is $\eta_1' \{\mathbf{K}_1 + L'\mathbf{K}_2\}$, implying that $\mathbf{K}_1 + L'\mathbf{K}_2$ is a minimal sufficient statistic of a smaller dimension, which is impossible. This establishes that $\eta_2(\eta_1)$ must be non-linear. \square

This corollary explains why in Example 3 the parameter is $\theta = (\mu, \tau^2 \mu^2)'$ and not, for example

$(\mu, \mu)'$. The latter case is studied in Example 6, a univariate normal sample with mean equal to the variance. The convenience of Proposition 1 is seen by generalizing Examples 3 and 6 to Examples 7 and 8, respectively. The first one is a univariate normal with coupled mean and variance, for which it is very difficult to find a function $g(k)$ that establishes incompleteness, with the use of the criterion is straightforward. Example 8 is a 2×2 contingency table with unconstrained parameterisation, leading to a complete sufficient statistic. Similar logic will be used in the next section, to illustrate a simple yet generic clustered-data setting. Details can be found in Appendix A.

5 Illustration: Clusters Following a Compound-Symmetry Model

First, consider univariate outcomes with random sample size. Example 9 is a univariate normal with unknown mean, unit variance, and random sample size. The sufficient statistic is then incomplete. In Example 10 this is extended to normal compound-symmetry vectors, where incompleteness evidently also applies. In Example 11 the same is seen to be true for the entire exponential family. In Example 12 the clusters are further allowed to be of variable size. The sufficient statistic is then still incomplete. In Example 13, this general result is applied to normal compound-symmetry data with clusters of unequal size, not allowing for a complete sufficient statistic. In Example 14, we allow for both random cluster sizes and a general exponential model formulation.

The use of Proposition 1 is trivial in this context. There are three model parameters, $\theta = (\mu, \sigma^2, \tau^2)'$, but the sufficient statistic is necessarily of higher dimension, as soon as there are at least two different cluster sizes. This route is easier than the explicit construction of a function (39). Even though this was still practicable, the computations for Example 4 are much more complex. This is because in Example 4 the stopping rule depends on the data, in contrast to in our most recent Examples 3–19, where the cluster sizes are completely random.

In summary, because $\eta(\theta)$ will generally be such that the dimension of η is higher than that of θ , Proposition 1 applies. The qualification ‘generally’ is needed, because there are obvious (trivial) counterexamples. In Example 13 the sufficient statistic for σ^2 is one-dimensional, as an

exception to the rule. When this would hold for all parameters, then completeness would hold. Such an example may be difficult to construct though. Another situation is when the cluster members are independent. Then the cluster sizes become irrelevant. If, in this special case, further $\sum_i N_i$, the overall sample size, would be constant, then such a clustered-data example reduces to a conventional univariate sample with fixed sample size, and completeness follows, establishing a counterexample. Apart from such pathological cases, virtually all practically relevant clustered data applications have incomplete minimal sufficient statistics.

6 Missing Data in Contingency Tables and Beyond

First consider the simple yet generic setting of missing data in contingency tables. We then turn to general missing data settings and end this section by bringing out commonality between seemingly disparate settings, considered earlier in this manuscript, that all lead to incomplete sufficient statistics.

In Example 15 a fully observed 2×1 contingency table is considered, which allows for a complete sufficient statistic. When data are partially missing (Example 16), this is no longer true. Example 16 and function (40) are reminiscent of Example 3, where function (17) exists because τ^2 is known. In spite of the similarity, there is an important difference as well: q is an *unknown* constant that nevertheless does not need to be estimated, because of ignorability. Admittedly, Examples 15–16 are very simple and therefore it is hard to see the generality of the result. Thus consider a 2×2 contingency table with supplemental margins as well (Example 17), then no complete sufficient statistic exists.

In the above examples, there is nothing particular about the use of contingency tables, nor about the parameterization used for the counts, leading to the following proposition.

Proposition 2. *Proposition 2 (Incomplete sufficient statistics with ignorable likelihood.) Let an exponential family model $f(\mathbf{Y}|\boldsymbol{\theta})$ admit a complete sufficient statistic when data are fully observed, then the same model does not admit a complete sufficient statistic under ignorable likelihood when data are partially missing.*

It is interesting to reflect upon the nature of this result. When data are partially missing, the data are effectively stratified, with one stratum grouping the fully observed trials and the other stratum the remaining trials. Still, the parameters of p -type (Example 16–17) describe both strata simultaneously. Because of ignorability, it is sensible to formulate a model where the parameter vector is of the same length as it would be when data were complete, but the stratification nevertheless implies that the length of the vector of sufficient statistics increases. This leads to the conclusion that this same phenomenon also occurs in other settings, including many non-missing-data settings. In Example 13, the strata are defined by the different cluster sizes occurring in the data. In that example, completeness could be restored by assuming that for every one of the cluster sizes n_ℓ occurring, there is a separate parameter vector $(\mu_\ell, \sigma_\ell^2, \tau_\ell^2)$, together with a multinomial vector (π_1, \dots, π_L) describing the probabilities with which the various cluster sizes occur.

Other examples can now be reconsidered. In Example 9, completeness would be established by estimating a separate parameter for each of the cluster sizes that can occur. The parameter would then be $(\mu_1, \dots, \mu_n; \pi_1, \dots, \pi_n)$. Obviously, in this particular example, this consideration is of theoretical interest only, for two reasons. First, the parameters μ_N may not be of direct scientific value. Second, from a given experiment, we can estimate only one of them, and which one it will be is random in itself. This is different in Example 13, where typically more than one cluster size is observed in a given experiment. The fact that the parameter depends on the cluster size is then not a theoretical consideration, but a well studied problem often indicated by the term informative cluster size (Chiang and Lee, 2008; Aerts *et al.*, 2011). It is different, too, in the missing-data examples: allowing for a different parameter in different strata (also called patterns of missingness), brings us to the so-called pattern-mixture model (Molenberghs and Kenward, 2007).

While an informal statement only, it is useful to see that many estimands do not allow complete sufficient statistics because the corresponding parameters are estimated from data where this same parameter describes two or more natural strata simultaneously. By ‘natural strata,’ we mean strata that lead to separate sufficient statistics for the same parameter, without the opportunity to combine these into a single one. Looking at this from a different angle, it

provides a basis for the following, existing, procedure. First, estimate separate copies of the parameter for every one of the strata. Second, combine these using appropriate weights. This procedure was studied by Hermans *et al.* (2016), based upon work by Molenberghs, Verbeke, and Iddi (2011).

7 Concluding Remarks

In this paper, building upon the work reported in Liu and Hall (1999), Liu *et al.* (2006), Molenberghs *et al.* (2014), and Milanzi *et al.* (2014, 2015), we have provided an easy-to-use criterion for incompleteness of minimal sufficient statistics in univariate and multivariate exponential family models. Earlier work has typically studied incompleteness directly by means of the definition. This either implies that the existence of a non-trivial zero-expectation function needs to be falsified, or that such a function needs to be constructed. Our result essentially requires checking the dimension of a minimal sufficient statistic relative to the length of the parameter vector. This turns the assessment of incompleteness into a feasible task, whereas the definition can be daunting to use and requires ad hoc construction of distributions of minimal sufficient statistics.

We have shown that clustered data designs with non-constant cluster sizes (random or otherwise) do not admit complete sufficient statistics. The term ‘clustered data’ has to be understood in the broadest sense; it encompasses longitudinal studies, multilevel designs, etc. On the one hand, longitudinal studies can have variable cluster sizes by design, while on the other, their cluster sizes can vary because of missing data.

The incompleteness of minimal sufficient statistics leads to the loss of some desirable properties, such as unbiasedness and optimality. But as shown in Molenberghs *et al.* (2014) and Milanzi *et al.* (2014, 2015), this does not need to be a serious problem in practice. For example, it is very well-known that, when data are missing, likelihood and Bayesian inferences can be based on the observed-data likelihood, without any correction for the variable cluster size, i.e., without any correction for the missing-data mechanism. Importantly, though, such methods cannot, in general, by default be claimed to be optimal, given that the Lehman-Scheffé

theorem (Casella and Berger, 2001) does not apply. The consequences for the case of random cluster sizes, in particular informative cluster sizes, are not widely understood. When cluster sizes follow a random mechanism (in the sense of missing at random), it is thus possible to simply use the observed-data likelihood without *ad hoc* corrections. However, one cannot claim that such an approach is ‘uniformly better’ than any of the dedicated corrections. Arguably, it is prudent to investigate candidate methods’ operational characteristics in settings relevant for the application at hand.

In the absence of complete sufficient statistics, some interesting philosophical issues appear. As discussed in Milanzi *et al.* (2015), some estimators will depend on the fact that more data could have been collected or that some data are available that, with certain probability, might not have been collected. They illustrated this using so-called *generalized sample averages* in sequential studies. When excluding such esoteric estimators, often only intuitively appealing estimators, such as the ordinary sample average, remain, even though there is no complete sufficient statistic, and in spite of some small-sample bias. Note, this shows great similarity with earlier work of Liu and Hall (1999) and Liu *et al.* (2006).

Our focus has been on characterizing incompleteness and, in particular, its consequences for point estimators. There obviously are important implications for hypothesis testing and interval estimation as well. An early reference is Anscombe (1949) and the topic, especially in the context of sequential designs, has received thorough treatment in Govindarajulu (1981), Barndorff-Nielsen and Cox (1984), and Barndorff-Nielsen and Cox (1994). More recently, members of the author team have studied the impact of incomplete sufficient statistics on estimation and hypothesis testing (Milanzi *et al.*, 2015), and the implications thereof for sequential designs (Milanzi *et al.*, 2014).

ACKNOWLEDGEMENTS

The authors acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy). The research leading to these results has also received funding from the European Seventh Framework programme FP7 2007–2013 under grant agreement Nr. 602552. We gratefully acknowledge support from the IWT-SBO ExaScience grant.

References

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.
- Aerts, M., Faes, C., Hens, N., Loquiha, O., and Molenberghs, G. (2011). Incomplete clustered data and non-ignorable cluster size. In: Conesa, D., Forte, A., López-Quílez, A. and Muñoz, F. (Eds.), *Proceedings of the 26th International Workshop on Statistical Modelling*, València, Spain, 35–40.
- Anscombe, F.J. (1949). Large-sample theory of sequential estimation. *Biometrika*, 36, 455–458.
- Armitage, P. (1975). *Sequential Medical Trials*. Oxford: Blackwell.
- Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics - Series B*, 53, 233-243.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review*, 52, 309–326.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. London: Chapman&Hall.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, 15, 377–380.
- Benhin, E., Rao, J.N.K., and Scott, A.J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, 92, 435–450.
- Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18,105-110.
- Boos D.D. and Stefanski L.A., (2013). *Essential Statistical Inference: Theory and Methods*. New-York: Springer.

- Brown L.D., (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics.
- Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.
- Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions – A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2, 6.
- Chiang, C-T., and Lee, K-Y. (2008). Efficient estimation methods for informative cluster size data. *Statistica Sinica*, 18, 121–133.
- Cong, X-J., Yin, G., and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*, 63, 663–672.
- Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77, 875–892.
- Geys, H., Molenberghs, G., and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, 62, 45-72.
- Govindarajulu, Z. (1981). *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory*. Colombus, OH: American Sciences Press.
- Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2016). Clusters with random size: maximum likelihood versus weighted estimation. *Submitted for publication*.
- Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001). Within-cluster resampling. *Biometrika*, 88, 1121–1134.
- 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.
- Keener, R.W. (2010). *Theoretical Statistics: Topics for a Core Course, Springer Texts in Statistics*. Springer 85–99.
- Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236–247.
- le Cessie, S. and van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, 43, 95-108.
- Lehmann, E.L., (1981). *Testing statistical hypothesis*. New York: John Wiley & Sons.
- Lehmann, E.L. and Stein, C. (1950). Completeness in the sequential case. *Annals of Mathematical Statistics*, 21, 376–385.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Liang, K., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, 54, 3-40.
- Lindsay, B. (1988), Composite likelihood methods. *Contemporary Mathematics*, 80, 220–239.
- 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.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*. London New York: Chapman and Hall.

- Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Tsiatis, A., Davidian, M., and Verbeke, G. (2015). Estimation after a group sequential trial. *Statistics in Biosciences*, 7, 187–205.
- Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Verbeke, G., Tsiatis, A.A., and Davidian, M. (2014). Properties of estimators in exponential family settings with observation-based stopping rules. *Journal of Biometrics & Biostatistics*, 7, 272.
- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.
- Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, 23, 11–41.
- Molenberghs, G., Verbeke, G., and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, 81, 892–901.
- Poularikas, A.D. and Seely, S. (2000). Laplace transforms. *The Transforms and Applications Handbook*. Boca Raton: Chapman & Hall/CRC Press
- Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, 75, 723–729.
- Rudin, W. (1974). *Real & Complex Analysis* (2nd ed.). New Delhi: McGraw Hill.
- Shao, T. (1999). *Mathematical Statistics* (2nd ed.). New York: Springer.
- 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.

- Van Garderen, K.J. (1997). Curved exponential models in econometrics. *Econometric Theory*, 1, 771–790.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16, 117–186.
- Wang, M., Kong, M., and Datta, S. (2011). Inference for marginal linear models for correlated longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research*, 20, 347–367.
- 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.
- Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59, 36–42.

A Tutorial on the Practical Use and Implication of Complete Sufficient Statistics

Lisa Hermans¹, Geert Molenberghs^{1,2}, Marc Aerts¹,
Michael G. Kenward³ and Geert Verbeke^{2,1}

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

² *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*

³ *Luton, United Kingdom*

E-mail: lisa.hermans@uhasselt.be

Supplementary Materials

A Examples

Example 1 (Univariate normal sample with known variance). *Let $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, with μ unknown and σ^2 known. Then $K_1 = \sum_{i=1}^n Y_i$ is a complete sufficient statistic for μ .*

Clearly, $K_1 \sim N(n\mu, n\sigma^2)$. Suppose that there is a function $g(k_1)$ such that $E\{g(k_1)\} = 0$ for all values of μ . Then

$$\int g(k_1) \frac{1}{\sqrt{n\sigma^2}\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \frac{(k_1 - n\mu)^2}{n\sigma^2}\right\} dk_1 = 0,$$

implying that

$$\int g(k_1) \exp\left\{-\frac{1}{2} \frac{k_1^2 \sigma^2}{n\sigma^4} + \frac{k_1}{\sigma^2} \mu\right\} d\left(\frac{k_1}{\sigma^2}\right) = 0.$$

With a simple change of variables, this can be written as

$$\int g(t\sigma^2) e^{-\frac{1}{2} \frac{t^2 \sigma^2}{n}} e^{t\mu} dt = \mathcal{L}\left\{g(t\sigma^2) e^{-\frac{1}{2} \frac{t^2 \sigma^2}{n}}\right\} = 0,$$

where $\mathcal{L}(\cdot)$ denotes the two-sided Laplace transform. This, in turn, implies that the argument must be equal to zero almost everywhere (a.e.). Because of the exponential factor, this forces $g(t\sigma^2) = 0$ a.e. Hence, $g(k_1) = 0$ a.e.

Example 2 (Univariate normal sample with unknown variance). *Let $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \dots, n$, with μ and σ^2 unknown. Then (K_1, K_2) with K_1 as in Example 1 and $K_2 = \sum_{i=1}^n Y_i^2$ is a complete sufficient statistic for (μ, σ^2) .*

The kernel of the log-likelihood, i.e., the terms of the log-likelihood that are functions of the parameters (McCullagh and Nelder, 1989), is

$$\begin{aligned}\ell &= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \\ &= -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n y_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n y_i - \frac{n\mu^2}{2\sigma^2}.\end{aligned}\tag{1}$$

The sufficient statistic (K_1, K_2) immediately follows. K_1 is normally distributed as in Example 1 and K_2 has a non-central chi-squared distribution. K_1 and K_2 are independent. Assume a function $g(k_1, k_2)$ with zero expectation for all values of the pair (μ, σ^2) .

Even though we have a bivariate statistic, we can start from the derivations in Example 1. Write the kernel of the density of K_m ($m = 1, 2$) as $h_m(k_m) \exp(\theta_m k_m)$, then the condition on $g(k_1, k_2)$ is:

$$0 = \int \int g(k_1, k_2) h_1(k_1) h_2(k_2) \exp(\theta_1 k_1 + \theta_2 k_2) dk_1 dk_2 \tag{2}$$

$$= \int dk_2 h_2(k_2) \exp(\theta_2 k_2) \int dk_1 g(k_1, k_2) h_1(k_1) \exp(\theta_1 k_1) \tag{3}$$

$$= \int dk_2 h_2(k_2) \exp(\theta_2 k_2) \mathcal{L}_{\theta_1} \{g(k_1, k_2) h_1(k_1)\} \tag{4}$$

$$= \mathcal{L}_{\theta_2} [h_2(k_2) \mathcal{L}_{\theta_1} \{g(k_1, k_2) h_1(k_1)\}], \tag{5}$$

where \mathcal{L}_{θ_1} is a two-sided and \mathcal{L}_{θ_2} a one-sided Laplace transform. This implies that $h_2(k_2) \mathcal{L}_{\theta_1} \{g(k_1, k_2) h_1(k_1)\} = 0$ a.e. and thus, because $h_2(k_2) > 0$ over the support, that $\mathcal{L}_{\theta_1} \{g(k_1, k_2) h_1(k_1)\} = 0$ a.e. This, in turn, implies that $g(k_1, k_2) h_1(k_1) = 0$ a.e. For a

reason similar to that used above, it follows that $g(k_1, k_2) = 0$ a.e. Note that a two-sided, or bilateral, Laplace transform is unique only, i.e., one-to-one onto its inverse, when not only the function but also the region of convergence is specified (Poularikas and Seely, 2000). However, in our case, because of the use of exponential family functions, the region of convergence is not restricted, hence this subtle issue does not apply here. In fact, an unrestricted region of convergence is a regularity condition: it is violated, for example, in the case of a deterministic stopping rule, but not when a stochastic stopping rule is used.

This derivation is quite general. Clearly, the argument can be extended to a vector of arbitrary length. Note however that we have used the fact that K_1 and K_2 are independent. While this is true for the mean and the variance related sufficient statistics for normal samples, it is not true in general. However, the extension to dependent sufficient statistics is almost trivial: we can replace $h_1(k_1)$ by $h_1(k_1|k_2)$ in (2)–(5). Furthermore, a univariate version of this argument generalizes Example 1.

The multivariate extension of this argument will be used in Section 4. It is not true however, that such an argument can cover all situations, for example, for the sequential trial case. A simple but instructive counterexample is provided next.

Example 3 (Univariate normal sample with coupled mean and variance). *Let $Y_i \sim N(\mu, \tau^2 \mu^2)$, $i = 1, \dots, n$, with μ unknown. In the case that τ^2 is known, there is no complete sufficient statistic for μ . On the other hand, when τ^2 is unknown, there is a complete sufficient statistic for (μ, τ^2) .*

The kernel of the log-likelihood immediately follows from (1), upon equating $\sigma^2 = \tau^2 \mu^2$:

$$\ell = -\frac{n}{2}(\ln \tau^2 + 2 \ln \mu) - \frac{1}{2\tau^2 \mu^2} \sum_{i=1}^n y_i^2 + \frac{1}{\tau^2 \mu} \sum_{i=1}^n y_i - \frac{n}{2\tau^2}. \quad (6)$$

Assume τ^2 known and consider the function

$$g(k_1, k_2) = \frac{k_1^2}{\tau^2 + n} - \frac{k_2}{\tau^2 + 1}. \quad (7)$$

Because $E(K_1^2) = n^2 \mu^2 + n \tau^2 \mu^2 = n \mu^2 (\tau^2 + n)$ and $E(K_2) = n \mu^2 (\tau^2 + 1)$, it readily follows

that the expectation of $g(K_1, K_2)$ is zero, while the function is non-trivial. Function $g(k_1, k_2)$ satisfies the definition of incompleteness only because τ^2 is a known constant.

The score equation for (6) can be written as:

$$n\mu^2 + \frac{K_1\mu}{\tau^2} - \frac{K_2}{\tau^2} = 0,$$

with solution

$$\hat{\mu} = \frac{-K_1 \pm \sqrt{K_1^2 + 4n\tau^2}}{2n\tau^2}.$$

Clearly, $\hat{\mu} + g(K_1, K_2)$ would provide another estimator with the same expectation, for every non-trivial function $g(k_1, k_2)$ with expectation zero. The derivations above show that this type of function exists. Adding such a function comes down to reweighing the amount of information taken from K_1 relative to that from K_2 .

This counterexample is interesting because, at first sight, it is close to Examples 1 and 2. However, in both of these earlier examples, the sufficient statistic and the parameter are of the same dimension, while here, the statistic is by necessity two-dimensional. If it is restricted to either K_1 or K_2 , then it is no longer sufficient.

But when τ^2 is unknown, the sufficient statistic and the parameter are again of the same dimension as in Example 2. The score for τ^2 leads to:

$$\tau^2 = \frac{2}{n} \left(\frac{K_2}{2\mu^2} - \frac{K_1}{\mu} + \frac{n}{2} \right)$$

and to solutions:

$$\hat{\mu} = \frac{K_1}{n}, \hat{\tau}^2 = \frac{nK_2}{K_1^2} - 1.$$

This example, with τ^2 known, is similar to the sequential-trial case where the sufficient statistic consists not only of the data collected, but also of the sample size realized, i.e., a one-dimensional parameter needs a two-dimensional sufficient statistic. The following example sets this out in some generality. It is based on developments in Milanzi *et al.* (2015). In this, a group sequential trial was considered with an arbitrary number of looks L and exponential

family distributed outcomes. It generalizes the results of Milanzi *et al.* (2014), who only considered a trial with two possible sample sizes, n and $2n$.

Example 4 (Sequential trial with stochastic stopping rule). *Consider a sequential trial with L pre-specified looks, with sample sizes $n_1 < n_2 < \dots < n_L$. Assume that there are n_j i.i.d. observations Y_1, \dots, Y_{n_j} , from the j th look that follow an exponential family distribution with density*

$$f_\theta(y) = h(y) \exp \{ \theta y - a(\theta) \}, \quad (8)$$

for θ the natural parameter, $a(\theta)$ the mean generating function, and $h(y)$ the normalizing constant. There is no complete sufficient statistic for the mean μ or, equivalently, for the natural parameter θ .

Subsequent developments are based on a generic data-dependent stochastic stopping rule, which we write as:

$$\pi(N = n_j | k_{n_j}) = F(k_{n_j} | \psi) = F(k_{n_j}), \quad (9)$$

where $k_{n_j} = \sum_{i=1}^{n_j} y_i$ is a realisation from an exponential family density:

$$f_{n_j}(k) = h_{n_j}(k) \exp \{ \theta k_{n_j} - n_j a(\theta) \}. \quad (10)$$

While we do not need to provide an explicit expression for the stopping rule at this point, as our developments apply to a broad class, it is useful to note that Milanzi *et al.* (2014) studied in detail the behaviour of stopping rules taking the form $F(\alpha + \beta k_{n_j} / n_j^m)$, for some power m and some cumulative distribution function $F(\cdot)$. Our inferential target is the parameter θ , or a function thereof.

In a sequential setting, a convenient minimal sufficient statistic is (K_3, N) , with $K_3 = \sum_{i=1}^N Y_i$. Following the developments in the above papers, the joint distribution for (K_3, N) is:

$$p(K_3, N) = f_0(K_3, N) F(K_N), \quad (11)$$

$$f_0(k_{n_1}, n_1) = f_{n_1}(k_{n_1}), \quad (12)$$

$$f_0(k_{n_j}, n_j) = \int f_0(k_{n_{j-1}}, n_{j-1}) f_{n_j - n_{j-1}}(k_{n_j} - k_{n_{j-1}}) [1 - F(k_{n_{j-1}})] dk_{n_{j-1}}. \quad (13)$$

If (K_3, N) were complete, then there would exist a function $g(K_3, N)$ such that $E[g(K_3, N)] = 0$ if and only if $g(K_3, N) = 0$, implying that

$$\begin{aligned} 0 = & \int g(k_{n_1}, n_1) f_{n_1}(k_{n_1}) F(k_{n_1}) dk_{n_1} + \sum_{j=2}^{L-2} \int g(k_{n_j}, n_j) H(k_{n_j}) F(k_{n_j}) dk_{n_j} \\ & + \int g(k_{n_L}, n_L) H(k_{n_L}) F(k_{n_L}) dk_{n_L}, \end{aligned} \quad (14)$$

with

$$H(k_{n_j}) = \underbrace{\int \dots \int}_{j-1} f_0(k_{n_{j-1}}, n_{j-1}) f_{n_j - n_{j-1}}(k_{n_j} - k_{n_{j-1}}) [1 - F(k_{n_{j-1}})] dk_{n_1} \dots dk_{n_{j-1}}.$$

Substituting the general exponential form (10) into (14), and applying properties of exponential family probability distributions, gives

$$\begin{aligned} 0 = & \int h_{n_L - n_1} e^{(\theta k_{n_1})} \int g(k_{n_1}, n_1) F(k_{n_1}) h_{n_1}(k_{n_1}) \exp(\theta k_{n_1}) dk_{n_1} \\ & + \sum_{j=2}^{L-2} \int h_{n_L - n_j} e^{(\theta k_{n_j})} \int g(k_{n_j}, n_j) \tilde{H}(k_{n_j}) \exp(\theta k_{n_j} - n_j) F(k_{n_j}) dk_{n_j} \\ & + \int g(k_{n_L}, n_L) \tilde{H}(k_{n_L}) \exp(\theta k_{n_L}) F(k_{n_L}) dk_{n_L}, \end{aligned} \quad (15)$$

where

$$\tilde{H}(k_{n_j}) = \left[\underbrace{\int \dots \int}_{j-1} \prod_{i=1}^{j-1} h_{n_1}(k_{n_1}) h_{n_{i+1} - n_i}(k_{n_{i+1}} - k_{n_i}) [1 - F(k_{n_i})] dk_{n_1} \dots dk_{n_{j-1}} \right].$$

Upon noting that the right hand side is a convolution and making use of properties of linearity and uniqueness of the Laplace transform it can be shown that:

$$\begin{aligned} g(k_{n_L}, n_L) \tilde{H}(k_{n_L}) &= - \sum_{j=1}^{L-1} \int g(z_j, n_j) \tilde{H}(z_j) F(z_j) dz_j, \\ g(k_{n_L}, n_L) &= \frac{\sum_{j=1}^{L-1} \int g(z_j, n_j) \tilde{H}(z_j) F(z_j) dz_j}{\tilde{H}(k_{n_L})}. \end{aligned}$$

Note that the Laplace transform is unique in both the unilateral as well as the bilateral case. In the unilateral case, this property is straightforward. In the bilateral case, the additional requirement needs to be added that this uniqueness holds over the region of absolute convergence. As mentioned earlier, this region of convergence is not restricted as a stochastic stopping rule is applied. Assigning, for example, arbitrary constants to $g(n_1, k_{n_1}), \dots, g(n_{L-1}, k_{n_{L-1}})$, a value can be found for $g(n_L, k_{n_L}) \neq 0$, contradicting the requirement for (K_3, N) to be complete, hence establishing incompleteness. From applying the Lehmann-Scheffé theorem, no best mean-unbiased estimator is guaranteed to exist. The practical consequence of this is that even estimators as simple as a sample average need careful consideration and comparison with alternatives. To do this, we embed the sample average in a broader class of linear estimator, and also study it from a likelihood perspective.

Consider the special case of $L = 2$, $n_1 = n$, and $n_2 = 2n$, a normally distributed endpoint with mean μ and variance 1, and probit probability of stopping after the first look equal to $\Phi(\alpha + \beta k/n)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Then, following Molenberghs *et al.* (2014), incompleteness is established by constructing a non-trivial function $g(K_3, N)$ (where K_3 is the sample sum and N is the realized sample size, i.e., N can take values n and $2n$), satisfying for all μ :

$$g(k, 2n) \cdot p_0(2n, k) = - \int \phi_n(k - z) \cdot g(z, n) \cdot \phi_n(z) \cdot \Phi\left(\alpha + \frac{\beta}{n}z\right) dz, \quad (16)$$

where $\phi(\cdot)$ is the standard normal density. Molenberghs *et al.* (2014) gave two examples of such a function, one of which being:

$$g(k, n) = \frac{\lambda}{\Phi\left(\alpha + \frac{\beta}{n}k\right)}, \quad (17)$$

$$g(k, 2n) = -\frac{\lambda}{1 - \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n + \beta^2}{2n}}}\right)}, \quad (18)$$

with $\lambda \neq 0$ an arbitrary constant.

Example 5 (Known parameter). *Consider the bivariate case studied in Example 2 but now such that θ_2 is known.*

The requirement for an expectation-zero function is:

$$\int dk_2 h_2(k_2) e^{-A(\theta_2)} e^{\theta_2 k_2} \int dk_1 g(k_1, k_2) h_1(k_1|k_2) e^{-A(\theta_1, k_2)} e^{\theta_1 k_1} = 0. \quad (19)$$

Choose $g(k_1, k_2) = g(k_2)$. Condition (19) becomes:

$$\begin{aligned} & e^{-A(\theta_2)} \int dk_2 h_2(k_2) e^{\theta_2 k_2} g(k_2) \int dk_1 h_1(k_1|k_2) e^{-A(\theta_1, k_2)} e^{\theta_1 k_1} \\ &= e^{-A(\theta_2)} \int dk_2 h_2(k_2) e^{\theta_2 k_2} g(k_2) = 0. \end{aligned} \quad (20)$$

Hence, we merely need to satisfy:

$$\int g(k_2) h_2(k_2) e^{\theta_2 k_2} dk_2 = 0. \quad (21)$$

Importantly, because θ_2 is known, the left hand side of (21) is not a Laplace transform. Interpreting (21) as an inner product, we need only find a function $g(k_2) \perp h_2(k_2) e^{\theta_2 k_2}$, which is straightforward.

Example 6 (Univariate normal sample with identical mean and variance). *Let $Y_i \sim N(\mu, \mu)$, $i = 1, \dots, n$. Then $K_2 = \sum_{i=1}^n Y_i^2$ is a complete sufficient statistic for μ .*

This example is surprisingly different from Example 3, because now the kernel of the log-likelihood is:

$$\ell = -\frac{n}{2} \ln \mu - \frac{1}{2\mu} \sum_{i=1}^n y_i^2 - \frac{n\mu}{2},$$

so K_1 disappears. We clearly have a scalar sufficient statistic, and completeness is trivial.

Note that the score equation takes the simple form

$$\mu^2 + \mu = \frac{K_2}{n},$$

leading to the maximum likelihood estimator:

$$\hat{\mu} = \frac{\sqrt{4K_2/n + 1} - 1}{2}.$$

In Example 4, the conditional likelihood accommodating both K_3 and N has a non-linear correction term relative to the ordinary least squares solution to the likelihood equations in the standard case of a fixed sample size (Molenberghs *et al.*, 2014; Milanzi *et al.*, 2014, 2015).

Example 7 (Univariate normal sample with general coupling of mean and variance). *Let $Y_i \sim N(\mu, \mu^{2\lambda})$, $i = 1, \dots, n$. Then there is a complete sufficient statistic for μ only for $\lambda = 0$ or $\lambda = 1/2$.*

When $\lambda = 0$, Example 1 is recovered. Example 6 follows for $\lambda = 1/2$. In all other cases, the sufficient statistic is bivariate, which follows from the kernel of the log-likelihood:

$$\ell(\mu) \propto -n\lambda \ln \mu - \frac{K_2}{2\mu^{2\lambda}} + \frac{K_1}{\mu^{2\lambda-1}} - \frac{n}{2\mu^{2\lambda-2}}.$$

Given that $K_1 \sim N(n\mu, n\mu^{2\lambda})$ and $K_2 \sim \chi_{n\mu^{2\lambda}}^2$, it follows that $E(K_1) = n\mu$, $E(K_2) = 2n\mu^{2\lambda}$, and $E(K_1^2) = n^2\mu^2 + n\mu^{2\lambda}$. Consider a function

$$g(k_1, k_2) = \alpha k_1 + \beta k_1^2 + \gamma k_2. \quad (22)$$

The expectation is

$$E \{g(K_1, K_2)\} = \alpha n\mu + \beta n^2\mu^2 + (\beta n + 2\gamma n)\mu^{2\lambda}.$$

When $\lambda = 1$ every choice $\gamma = -\beta(n+1)/2$ produces a non-zero function with zero expectation. For $\lambda \neq 1$, in addition to being different from 0 and 1/2 as well of course, there is no non-trivial solution. However, from Proposition 1, we know that for all $\lambda \neq 0, 1/2$, the sufficient statistic is incomplete. So it is seen that it is not because the posited function (22) fails to provide a counterexample that there exists none. We now know there are such functions, but the proposition obviates the need to explicitly construct one.

Next, we provide an additional example, using a contingency table.

Example 8 (Bivariate contingency table). *Consider an $M_1 \times M_2$ contingency table with conditional row probabilities $\varphi(k_1|k_2)$ and marginal column probabilities $\pi(k_2)$.*

First, assume that all probabilities are unknown and to be estimated. Assume that there is a function $g(k_1, k_2)$ with zero expectation. Then

$$\sum_{k_1=1}^{M_1} \sum_{k_2=1}^{M_2} g(k_1, k_2) \varphi(k_1|k_2) \pi(k_2) = 0, \quad (23)$$

with sum constraints on the parameters: $\sum_{k_2=1}^{M_2} \pi(k_2) = 1$ and $\sum_{k_1=1}^{M_1} \varphi(k_1|k_2) = 1$, for every value of k_2 . Because (23) should hold for all values of the parameters, $g(k_1, k_2) = 0$ follows immediately from algebraic results on polynomials.

Second, assume that $\pi(k_2)$ is given and choose $g(k_1, k_2) = g(k_2)$. Then, (23) simplifies to

$$\sum_{k_1=1}^{M_1} \sum_{k_2=1}^{M_2} g(k_2) \varphi(k_1|k_2) \pi(k_2) = \sum_{k_2=1}^{M_2} g(k_2) \pi(k_2) = 0.$$

Because the vector π is given, we merely need a set of constants g such that $g \perp \pi$.

Example 9 (Univariate outcomes with random sample size). *Consider $Y_i \sim N(\mu, 1)$, with sample size N , with $1 \leq N \leq n$ and the probability of realizing sample size N equal to π_N . The sufficient statistic for μ is incomplete.*

The sufficient statistic is (K_3, N) with $K_3 = \sum_{i=1}^N Y_i$ and N the usual sample size. Assume that all $\pi_N > 0$, for $N = 1, \dots, n$; this simplifies the calculations without loss of generality. Choose a function $g(k, N) = a_N$. It then follows that

$$\begin{aligned} E \{g(K_3, N)\} &= \int \sum_{N=1}^n g(k, N) \pi_N \phi(k; N\mu, N) dk \\ &= \sum_{N=1}^n a_N \pi_N \int \phi(k; N\mu, N) dk \\ &= \sum_{N=1}^n a_N \pi_N. \end{aligned}$$

This expectation equals zero if a vector $\mathbf{a} \perp \boldsymbol{\pi}$ is chosen. Choosing (a_1, \dots, a_{n-1}) freely, then

$$a_n = -\frac{1}{\pi_n} \sum_{N=1}^{n-1} a_N \pi_N$$

satisfies the requirement. In the next example, we consider clustering between the outcomes.

Example 10 (Correlated outcomes with compound-symmetry structure and random sample size). *The setting is similar to that of Example 9, except that the vector $\mathbf{Y} \sim N(\mu \mathbf{1}_N, \sigma^2 I_N + \tau^2 J_N)$, with $\mathbf{1}_N$ a vector of ones of length N , I_N the N -dimensional identity matrix, and J_N an $N \times N$ matrix of ones. The sufficient statistic for (μ, σ^2, τ^2) is incomplete.*

The sufficient statistic is $(K_3 = \sum_{i=1}^N Y_i, K_4 = \mathbf{Y}'\mathbf{Y}, K_5 = \mathbf{Y}'J_N\mathbf{Y}, N)$, as will be clear from Example 13. By choosing a function $g(k_1, k_2, k_3, N) = a_N$ the same solution $\mathbf{a} \perp \boldsymbol{\pi}$ follows. This result does not depend in any way on this particular normality assumption, as can be formalized in the next example.

Example 11 (Vector-valued data and parameter, with completely random sample size). *Assume an exponential family structure $f(\mathbf{k}, N) = f_N(\mathbf{k})\pi(N|\mathbf{k}) \stackrel{\text{notation}}{=} f_N(\mathbf{k})\pi_N(\mathbf{k})$. The sufficient statistic is incomplete.*

Choose $g(\mathbf{k}, N) = g_N(\mathbf{k}) = a_N/\pi_N(\mathbf{k})$ for $\pi_N(\mathbf{k}) \neq 0$ and 0 otherwise. Then

$$E\{g(K_3, N)\} = \sum_{N=1}^n \int f_N(\mathbf{k})\pi_N(\mathbf{k})g_N(\mathbf{k})d\mathbf{k} = \sum_{N=1}^n a_N \int f_N(\mathbf{k})d\mathbf{k} = \sum_{N=1}^n a_N = 0$$

for any zero-sum sequence.

Of course, by using the term clustered data we do imply that N clusters of sizes N_i ($i = 1, \dots, m$) are sampled. We have not considered this level of generality yet. Example 11 will be generalized next.

Example 12 (N clusters of completely random size). *Consider N clusters of sizes N_i ($i = 1, \dots, N$), with sufficient statistics $[\mathbf{K} = \mathbf{K}\{(\mathbf{Y}_i)\}; \mathbf{N} = \mathbf{N}\{(N_i)\}]$. The sufficient statistic is incomplete.*

This result has the same form as in the previous example, with $g_N(\mathbf{k}) = a_N/\pi_N(\mathbf{k})$ this time, and

$$\sum_N a_N = 0.$$

Example 13 (Compound-symmetry clusters of random size). *Consider clustered data $\mathbf{Y}_i \sim N(\mu \mathbf{1}_{N_i}, \sigma^2 I_{N_i} + \tau^2 J_{N_i})$, for $i = 1, \dots, N$. The sufficient statistic for (μ, σ^2, τ^2) is incomplete.*

The terms in the log-likelihood that are data-dependent, and hence produce the sufficient statistic, follow from

$$\begin{aligned} & \sum_{i=1}^N -\frac{1}{2}(\mathbf{Y}_i - \mu \mathbf{1}_{N_i})'(\sigma^2 I_{N_i} + \tau^2 J_{N_i})^{-1}(\mathbf{Y}_i - \mu \mathbf{1}_{N_i}) \\ &= \sum_{i=1}^N -\frac{1}{2}(\mathbf{Y}_i - \mu \mathbf{1}_{N_i})' \left(I_{N_i} - \frac{\tau^2}{\sigma^2 + N_i \tau^2} J_{N_i} \right) (\mathbf{Y}_i - \mu \mathbf{1}_{N_i}) \\ &= \sum_{i=1}^N \frac{\mu}{\sigma^2 + N_i \tau^2} \left(\sum_{j=1}^{N_i} Y_{ij} \right) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^N \sum_{j=1}^{N_i} Y_{ij}^2 \right) \\ & \quad + \sum_{i=1}^N \frac{\tau^2}{2\sigma^2(\sigma^2 + N_i \tau^2)} \left(\sum_{j=1}^{N_i} Y_{ij} \right)^2. \end{aligned} \tag{24}$$

The three terms in (24) are qualitatively different. Indeed, the middle one corresponds to a single sufficient statistic, the sum of all squares across clusters, while the first and last split into as many sufficient statistics as there are unique cluster sizes. To properly formalize this, assume that there are L different cluster sizes, and that there are c_ℓ clusters among the data of size n_ℓ . Evidently, $m = \sum_{\ell=1}^L c_\ell$. Based on (24) and the multiplicity of the cluster sizes,

the sufficient statistics are:

$$S_{1\ell} = \sum_{i=1}^{c_\ell} \sum_{j=1}^{n_\ell} Y_{ij}^{(\ell)}, \quad (25)$$

$$S_2 = \sum_{\ell=1}^L \sum_{i=1}^{c_\ell} \sum_{j=1}^{n_\ell} \left(Y_{ij}^{(\ell)} \right)^2, \quad (26)$$

$$S_{3\ell} = \sum_{i=1}^{c_\ell} \left(\sum_{j=1}^{n_\ell} Y_{ij}^{(\ell)} \right)^2, \quad (27)$$

$$S_{4\ell} = c_\ell, \quad (28)$$

where the superscript (ℓ) is used to indicate that the summation is restricted to data from clusters of size n_ℓ . The conditional and marginal expectations of (25)–(28) are:

$$\begin{aligned} E(S_{1\ell}|c_\ell) &= c_\ell n_\ell \mu, \\ E(S_{1\ell}) &= m \mu \pi_\ell n_\ell, \\ E(S_2|c_\ell) &= \sum_{\ell=1}^L c_\ell n_\ell (\sigma^2 + \tau^2 + \mu^2), \\ E(S_2) &= N(\sigma^2 + \tau^2 + \mu^2) \sum_{\ell=1}^L \pi_\ell n_\ell, \\ E(S_{3\ell}|c_\ell) &= c_\ell \{ n_\ell (\sigma^2 + \tau^2 + \mu^2) + n_\ell (n_\ell - 1) (\tau^2 + \mu^2) \}, \\ E(S_{3\ell}) &= m \pi_\ell n_\ell \{ (\sigma^2 + \tau^2 + \mu^2) + (n_\ell - 1) (\tau^2 + \mu^2) \}, \\ E(S_{4\ell}) &= m \pi_\ell. \end{aligned}$$

Group all sufficient statistics into S and define a function

$$g(s) = \sum_{\ell=1}^L \lambda_\ell \frac{S_{1\ell}}{S_{4\ell}}. \quad (29)$$

Then,

$$E \{ g(S|S_4) \} = \sum_{\ell=1}^L \lambda_\ell \frac{E(S_{1\ell}|S_{4\ell})}{S_{4\ell}} = \mu \sum_{\ell=1}^L \lambda_\ell n_\ell,$$

and hence

$$E \{g(S)\} = \mu \sum_{\ell=1}^L \lambda_{\ell} n_{\ell}.$$

Once again, every solution $\lambda \perp \mathbf{n}$, where $\mathbf{n} = (n_1, \dots, n_L)'$, provides a counterexample, establishing incompleteness.

Example 14 (General clustered-data setting with random cluster sizes). *Consider clustered data \mathbf{Y}_i of size N_i , for $i = 1, \dots, N$, following an exponential family with data- and cluster-size components $f(\mathbf{y}_i|\boldsymbol{\theta}, N_i)$ and $f(N_i|\boldsymbol{\psi})$. Whenever N_i can take more than one value, the sufficient statistic for $\boldsymbol{\theta}$ is generally incomplete.*

Example 15 (Fully observed 2×1 contingency table). *Consider a binomial experiment based on a binary variable Y_i taking values 1 and 2, with n trials and parameter p ($i = 1, \dots, n$). Denote the number of 1s and 2s by Z_{21} and Z_{22} , respectively, such that $Z_{21} + Z_{22} = n$. The sufficient statistic for p is complete.*

(The first of the double index is redundant in this example, but is needed in the following one.) Because of the sum constraint, the sufficient statistic is Z_{21} (or Z_{22}), and the MLE is $\hat{p} = Z_{21}/n$. The result is obvious. Now turn to the same setting where not all observations are made.

Example 16 (Partially missing 2×1 contingency table). *Consider a binomial experiment based on a binary variable Y_i taking values 1 and 2, with n trials and parameter p ($i = 1, \dots, n$). Denote the number of 1s and 2s by Z_{21} and Z_{22} , respectively, and let the number of trials with unobserved outcome be Z_1 . Then, $Z_{21} + Z_{22} + Z_1 = n$. Assume that the missing data are missing at random. The sufficient statistic is incomplete if ignorable likelihood is used.*

In the above, missing at random means that the missing data mechanism does not depend on unobserved information, given observed information. Under missingness at random, mild regularity conditions, and drawing likelihood inferences, it is well-known that the missing-data mechanism can be ignored. For details, see Little and Rubin (2002).

Let $R_i = 1$ if Y_i is observed and $R_i = 0$ otherwise. Further, let $q = P(Y_i = 1)$. Full likelihood means that p and q are both estimated from the data. It is easy to show that $\hat{p} = Z_{21}/(Z_{21} + Z_{22})$ and $\hat{q} = (Z_{21} + Z_{22})/n$. When both parameters are estimated, the sufficient statistic (because of the sum constraint) and the parameter vector are both two-dimensional, establishing completeness. However, under missingness at random the likelihood factors into a factor containing p only and a factor with only q . It is then common practice to ignore the factor containing q and to restrict efforts to estimation of p . This leads to the same estimator for p . The sufficient statistic remains two-dimensional: both Z_{21} and Z_{22} , because their sum is random as well, unlike in the non-missing-data case. It is then easy to construct a function $g(z_{21}, z_{22})$, such that $E[g(Z_{21}, Z_{22})] = 0$ for every value of p :

$$g(z_{21}, z_{22}) = \frac{Z_{21} + Z_{22}}{q} - \frac{Z_{21}}{1 - q}. \quad (30)$$

Example 17 (Partially missing 2×2 contingency table). *Consider a contingency table with supplemental margin, cross-classifying two binary outcomes (Y_{i1}, Y_{i2}) , $(i = 1, \dots, n)$, and with counts Z_{2jk} and Z_{1j} ($j, k = 1, 2$). Unless the supplemental margin is empty, the sufficient statistic for the response profile probabilities p_{jk} under ignorable likelihood is incomplete.*

Under ignorability, only the probabilities p_{jk} are estimated (subject to their sum being one), and not the missingness probabilities q_j , where q_j is the probability of observing the second outcome Y_{i2} for a subject with $Y_{i1} = j$. Because $E(Z_{2jk}) = np_{jk}q_j$ and $E(Z_{1j}) = np_{j+}(1 - q_j)$, where the $+$ sign instead of k indicates summation over k , it follows that the functions

$$E[g_j(Z_{2j1}, Z_{2j2})] = (1 - q_j)(Z_{2j1} + Z_{2j2}) - q_j Z_{1j},$$

($j = 1, 2$), have zero expectation.

Example 18 (Standard exponential distribution for continuous times). *Consider Y_i ($i = 1, \dots, n$) i.i.d. with exponential density $f(y_i) = \lambda e^{-\lambda y_i}$. The parameter is λ , the sufficient statistic K_1 is complete.*

The first derivative of the log-likelihood based on the above model is

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - K_1,$$

from which it clearly follows that the dimension of both parameter and minimal sufficient statistic are equal to one.

Example 19 (Standard Poisson distribution for count data). *Consider Y_i ($i = 1, \dots, n$) i.i.d. with Poisson probability $P(y_i) = \frac{1}{y_i!} \lambda^{y_i} e^{-\lambda}$. The parameter is λ , the sufficient statistic K_1 is complete.*

The first derivative of the log-likelihood is

$$\frac{\partial \ell}{\partial \lambda} = \frac{1}{\lambda} K_1 - n,$$

from which it follows also here that the dimension of both parameter and minimal sufficient statistic are equal to one.

Example 20 (Integrated exponential probabilities for counts). *Consider Y_i ($i = 1, \dots, n$) i.i.d. with probabilities following from integrating the exponential density between two subsequent integer values: $P(y_i) = e^{-\lambda y_i} (1 - e^{-\lambda})$. The parameter is λ , the sufficient statistic K_1 is complete.*

The first derivative of the log-likelihood is

$$\frac{\partial \ell}{\partial \lambda} = -K_1 + n \frac{e^{-\lambda}}{1 - e^{-\lambda}},$$

from which it follows once more that the dimension of both the parameter as well as the minimal sufficient statistics are equal to one.

Note that, while in Examples 18–19 the estimators for λ are equal to the sample average $\hat{\lambda} = \bar{Y} = K_1/n$, for Example 20 the estimator is

$$\hat{\lambda} = -\ln \left(\frac{\bar{Y}}{1 + \bar{Y}} \right).$$

Of course, this difference is inconsequential for the completeness result.

Example 21 (Integrated Weibull probabilities for counts). *Consider Y_i ($i = 1, \dots, n$) i.i.d. with probabilities following from integrating the Weibull density between two subsequent integer values:*

$$P(y_i) = e^{-\lambda y_i^\rho} - e^{-\lambda(y_i+1)^\rho}.$$

The parameter is (λ, ρ) , representing location and shape, but no reduction in statistics is possible, i.e., it consists of all individual values $(Y_i)_i$.

The log-likelihood derivatives are:

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda} &= \sum_{i=1}^n \frac{-e^{-\lambda y_i^\rho} y_i^\rho + e^{-\lambda(y_i+1)^\rho} (y_i+1)^\rho}{e^{-\lambda y_i^\rho} - e^{-\lambda(y_i+1)^\rho}}, \\ \frac{\partial \ell}{\partial \rho} &= \sum_{i=1}^n \frac{-e^{-\lambda y_i^\rho} y_i^\rho \ln(y_i) + e^{-\lambda(y_i+1)^\rho} (y_i+1)^\rho \ln(y_i+1)}{e^{-\lambda y_i^\rho} - e^{-\lambda(y_i+1)^\rho}}. \end{aligned}$$

Clearly, no dimension reduction of the data is possible: the parameter is two-dimensional, but the sufficient statistic is of length n . Upon noting that

$$E(y_i) = \sum_{n=0}^{+\infty} e^{-\lambda n^\rho} - 1 = \alpha,$$

(α is used for notational purposes) it follows that a function

$$g(y_1, \dots, y_n) = \sum_{i=1}^n \beta_i y_i,$$

has expectation

$$E[g(y_1, \dots, y_n)] = \alpha \sum_{i=1}^n \beta_i,$$

which is equal to zero for any zero-sum (contrast) vector $(\beta_1, \dots, \beta_n)'$.

References

- Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.
- Aerts, M., Faes, C., Hens, N., Loquiha, O., and Molenberghs, G. (2011). Incomplete clustered data and non-ignorable cluster size. In: Conesa, D., Forte, A., López-Quílez, A. and Muñoz, F. (Eds.), *Proceedings of the 26th International Workshop on Statistical Modelling*, València, Spain, 35–40.
- Anscombe, F.J. (1949). Large-sample theory of sequential estimation. *Biometrika*, 36, 455–458.
- Armitage, P. (1975). *Sequential Medical Trials*. Oxford: Blackwell.
- Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics - Series B*, 53, 233-243.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review*, 52, 309–326.
- Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*. London: Chapman&Hall.
- Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, 15, 377–380.
- Benhin, E., Rao, J.N.K., and Scott, A.J. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika*, 92, 435–450.
- Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, 18,105-110.
- Boos D.D. and Stefanski L.A., (2013). *Essential Statistical Inference: Theory and Methods*. New-York: Springer.

- Brown L.D., (1986). *Fundamentals of Statistical Exponential Families: with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics.
- Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.
- Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions – A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2, 6.
- Chiang, C-T., and Lee, K-Y. (2008). Efficient estimation methods for informative cluster size data. *Statistica Sinica*, 18, 121–133.
- Cong, X-J., Yin, G., and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*, 63, 663–672.
- Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77, 875–892.
- Geys, H., Molenberghs, G., and Lipsitz, S.R. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginal odds ratio models. *Journal of Statistical Computation and Simulation*, 62, 45-72.
- Govindarajulu, Z. (1981). *The Sequential Statistical Analysis of Hypothesis Testing, Point and Interval Estimation, and Decision Theory*. Colombus, OH: American Sciences Press.
- Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M.G., Van der Elst, W., Aerts, M., and Verbeke, G. (2016). Clusters with random size: maximum likelihood versus weighted estimation. *Submitted for publication*.
- Hoffman, E.B., Sen, P.K., and Weinberg, C.R. (2001). Within-cluster resampling. *Biometrika*, 88, 1121–1134.
- 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.
- Keener, R.W. (2010). *Theoretical Statistics: Topics for a Core Course, Springer Texts in Statistics*. Springer 85–99.
- Kenward, M.G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, 13, 236–247.
- le Cessie, S. and van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, 43, 95-108.
- Lehmann, E.L., (1981). *Testing statistical hypothesis*. New York: John Wiley & Sons.
- Lehmann, E.L. and Stein, C. (1950). Completeness in the sequential case. *Annals of Mathematical Statistics*, 21, 376–385.
- Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Liang, K., Zeger, S.L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B*, 54, 3-40.
- Lindsay, B. (1988), Composite likelihood methods. *Contemporary Mathematics*, 80, 220–239.
- 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.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
- McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*. London New York: Chapman and Hall.

- Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Tsiatis, A., Davidian, M., and Verbeke, G. (2015). Estimation after a group sequential trial. *Statistics in Biosciences*, 7, 187–205.
- Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Verbeke, G., Tsiatis, A.A., and Davidian, M. (2014). Properties of estimators in exponential family settings with observation-based stopping rules. *Journal of Biometrics & Biostatistics*, 7, 272.
- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley & Sons.
- Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2014). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, 23, 11–41.
- Molenberghs, G., Verbeke, G., and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, 81, 892–901.
- Poularikas, A.D. and Seely, S. (2000). Laplace transforms. *The Transforms and Applications Handbook*. Boca Raton: Chapman & Hall/CRC Press
- Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, 75, 723–729.
- Rudin, W. (1974). *Real & Complex Analysis* (2nd ed.). New Delhi: McGraw Hill.
- Shao, T. (1999). *Mathematical Statistics* (2nd ed.). New York: Springer.
- 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.

- Van Garderen, K.J. (1997). Curved exponential models in econometrics. *Econometric Theory*, 1, 771–790.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16, 117–186.
- Wang, M., Kong, M., and Datta, S. (2011). Inference for marginal linear models for correlated longitudinal data with potentially informative cluster sizes. *Statistical Methods in Medical Research*, 20, 347–367.
- 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.
- Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics*, 59, 36–42.