Supplementary materials for this article are available at http://pubs.amstat.org/toc/tas/63/4.

The Effective Sample Size and an Alternative Small-Sample Degrees-of-Freedom Method

Christel FAES, Geert MOLENBERGHS, Marc AERTS, Geert VERBEKE, and Michael G. KENWARD

Correlated data frequently arise in contexts such as, for example, repeated measures and meta-analysis. The amount of information in such data depends not only on the sample size, but also on the structure and strength of the correlations among observations from the same independent block. A general concept is discussed, the *effective sample size*, as a way of quantifying the amount of information in such data. It is defined as the sample size one would need in an independent sample to equal the amount of information in the actual correlated sample. This concept is widely applicable, for Gaussian data and beyond, and provides important insight. For example, it helps explain why fixed-effects and random-effects inferences of meta-analytic data can be so radically divergent. Further, we show that in some cases the amount of information is bounded, even when the number of measures per independent block approaches infinity. We use the method to devise a new denominator degrees-of-freedom method for fixed-effects testing. It is compared to the classical Satterthwaite and Kenward–Roger methods for performance and, more importantly, to enhance insight. A key feature of the proposed degrees-of-freedom method is that it, unlike the others, can be used for non-Gaussian data, too.

This article has supplementary material online.

KEY WORDS: Amount of information; Correlated data; Information limit; Mixed models; Small-sample inference.

1. INTRODUCTION

Information is an important concept in statistics, whether for parameter estimation, hypothesis testing, or other modes of inferences. It is well known (Cox and Hinkley 1974; Welsh 1996) that, in the context of a univariate Gaussian sample, the amount of information is represented by the sample size. The same is true for binary data, up to a variance-related factor. For time-to-event data, events rather than sample size is the central concept. However, as soon as one ventures away from independence, there is considerable uncertainty as to how to define the sample size, except in a number of well-understood cases, mostly in a multivariate normal setting (Johnson and Wichern 1992). At first sight, if several measurements are taken for each independent unit, one could take the number of individuals or the number of measurements as the sample size. Whereas the first approach underestimates the amount of information, the second one leads to overstatement. The reason is that the amount of information also depends on the correlation among the observations, not only in strength, but also in terms of the structure such correlations assume. We will see that there are important differences between, for example, compound-symmetric and first-order autoregressive structures.

Such considerations are important to assess the amount of information available. We argue that the concept of *effective sample size* (ess), which will be reviewed, revisited, and amplified in this article, can be seen as key to such understanding. For example, it aids understanding as to why there can be such large differences when, say, testing for treatment effect in a meta-analysis using a fixed-effects approach on the one hand and a random-effects approach on the other hand. Further, it is crucial to assess the amount of information available when conducting hypothesis testing in a finite-sample context. In a number of correlated contexts, especially with normally distributed outcomes, exact test statistics are known (Johnson and Wichern 1992). These include the well-known t-, F-, and U-tests. They are useful, but limited to such contexts as full multivariate normal with unstructured or compound-symmetry correlation structure. More complex settings abound, even when data are Gaussian, in repeated measures or otherwise hierarchical studies, when data are unbalanced in the sense of differential numbers of measures per subjects, perhaps not taken at a common set of time points, and/or with alternative covariance structures. It is for such contexts that Satterthwaite (1941) and Kenward and Roger (1997) numerator degrees-of-freedom methods have been considered. Also the work by Fay and Graubard (2001) for generalized estimating equations, covering, among others, the binary and count data cases, is noteworthy. Parallel work for non-Gaussian settings, such as the commonly encountered binary and count data settings, is virtually nonexistent and one often falls back on crudely using the residual number of degrees of freedom, defined as the number of measurements minus the number of parameters to be estimated. Alternatively, one then switches to an asymptotic method. The effective sample size

Christel Faes is Assistant Professor (E-mail: *christael.faes@uhasselt.be*) and Geert Molenberghs (E-mail: *geert.molenberghs@uhasselt.be*) and Marc Aerts (E-mail: *marc.aerts@uhasselt.be*) are Professors of Biostatistics, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium. Geert Verbeke is Professor of Biostatistics, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Katholieke Universiteit Leuven, Belgium (E-mail: *geert.verbeke@med.kuleuven.be*). Michael G. Kenward is Professor of Biostatistics, Medical Statistics Unit, London School of Hygiene and Tropical Medicine, U.K. All authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy). This work was partially supported by the Research Foundation—Flanders.

 DOI: 10.1198/tast.2009.08196

(ess) also leads to a degrees-of-freedom method, applicable in both Gaussian and non-Gaussian settings. We consider this a worthwhile contribution in addition to what is available in the literature.

The context for our illustrations will be correlated data, in particular mixed-effects models. Obviously, the ideas are also applicable to other contexts, such as stratified versus unstratified analysis, analysis of covariance versus t-tests with a baseline covariate, and cross-over trials (Jones and Kenward 2003) versus parallel-group design. In such contexts, the ess provides a way to contrast the information obtained under a more elaborate design with the simpler one. This underscores that the concept's appeal goes well beyond its use as an alternative small-sample degrees-of-freedom device.

This article is organized as follows. Notation and key models are introduced in Section 2. In Sections 4 and 5, the concepts of the effective sample size and the information limit are explained. The use of the effective sample size in the degrees-of-freedom context is discussed in Section 6, and the behavior of the various degrees-of-freedom methods is scrutinized and compared in Section 7. The case studies, introduced in Section 3, are analyzed in Section 8.

2. NOTATION AND MODELS

Assume an outcome Y_{ij} is measured on independent units $i = 1, \ldots, N$ with $j = 1, \ldots, n_i$ replications. Examples include n_i repeated measures on patient i in a clinical trial, or n_i patients in trial i of a meta-analysis. Group the outcomes in the n_i-dimensional vector $\mathbf{Y}_i$. A general linear mixed model decomposes $\mathbf{Y}_i$ as

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

(Verbeke and Molenberghs 2000), in which $\boldsymbol{\beta}$ is a vector of population-average regression coefficients called fixed effects, and where $\mathbf{b}_i$ is a vector of subject-specific regression coefficients. The $\mathbf{b}_i$ describe how the evolution of the ith subject deviates from the average evolution in the population. The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The random effects $\mathbf{b}_i$ and residual components $\boldsymbol{\varepsilon}_i$ are assumed to be independent with distributions $N(\mathbf{0}, D)$, and $N(\mathbf{0}, \Sigma_i)$, respectively. Thus, in summary,

$$\mathbf{Y}_i|\mathbf{b}_i \sim N(X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i, \Sigma_i), \qquad \mathbf{b}_i \sim N(\mathbf{0}, D). \tag{2}$$

The corresponding marginal model is $\mathbf{Y}_i \sim N(X_i\boldsymbol{\beta}, V_i = Z_iDZ_i' + \Sigma_i)$. For occasion j, Model (1) takes the form: $Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i + \varepsilon_{ij}$, in an obvious notation. We will frequently make use of the so-called random-intercepts version:

$$Y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + b_i + \varepsilon_{ij}, \tag{3}$$

with $b_i \sim N(0, \tau^2)$ and $\epsilon_{ij} \sim N(0, \sigma^2)$. The corresponding marginal model is $\mathbf{Y}_i \sim N(X_i\boldsymbol{\beta}, V_i = \sigma^2 I_{n_i} + \tau^2 J_{n_i \times n_i})$, where I_{n_i} and $J_{n_i \times n_i}$ are an $n_i \times n_i$ identity matrix and a matrix of ones, respectively.

3. APPLICATION

We motivate and illustrate this work using two different, generic settings. A third application is given in the Supplementary Materials.

3.1 Cancer of the Ovaries

Consider data from a meta-analysis of two large multicenter trials in advanced ovarian cancer (Ovarian Cancer Meta-Analysis Project 1991). The trials contain 411 and 382 patients, respectively. The survival times (in years) of individual patients are available in these trials. The endpoint of interest is the logarithm of survival, defined as time (in years) from randomization to death from any cause.

We consider a random-intercepts model (3) for Y_{ij}, the log-survival time of individual j in trial i. Our focus is on assessing the impact of the within-trial correlation; for this, we consider a simple model with the fixed-effects structure reduced to the overall mean log-survival time. It should be noted that this example is different in nature from a typical longitudinal study, because here only two trials contribute independent information. It will be shown that different estimation methods for the degrees of freedom may lead to major differences in the resulting p-values.

3.2 Rats Data

The data from this example resulted from a randomized longitudinal experiment (Verdonck et al. 1998), in which 50 male Wistar rats were randomized to either a control group or one of the two treatment groups, where treatment consisted of a low or high dose of the testosterone inhibitor Decapeptyl. The treatment started at the age of 45 days, and measurements were taken every 10 days, starting at the age of 50 days. Of interest was skull height, measured as the distance (in pixels) between two well-defined points on X-ray pictures taken under anesthesia. Some rats have incomplete follow-up because they did not survive anesthesia.

Let Y_{ij} denote the response taken at time t_j, for rat i. Similarly to the work of Verbeke and Molenberghs (2000), we model subject-specific profiles as linear functions of $t_j = \ln(1 + (\text{Age}_j - 45)/10)$, using (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta_0 + \beta_1 t_j$. Here, β_0 is the average response at the time of randomization, whereas β_1 is the average slope in the three different treatment groups. We are interested in $H_0 : \beta_1 = 0$, assessing the linear trend over time.

4. THE EFFECTIVE SAMPLE SIZE

In this section, the general concept of the effective sample size is reviewed and revisited. The idea is simple and appealing at the same time, and has been considered in a variety of contexts, such as climatic applications (Thiebaux and Zwiers 1984), complex survey methods (Skinner, Holt, and Smith 1989), and in a spatial setting (Cressie 1991).

We will set out by considering the Gaussian case. Assume model (1). The fixed-effects parameter $\boldsymbol{\beta}$ can be estimated as (Laird and Ware 1982):

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i\right)^{-1} \sum_{i=1}^{N} X_i'V_i^{-1}\mathbf{Y}_i.$$

This is an unbiased estimate for $\boldsymbol{\beta}$ if the mean of the response is correctly specified, even if the variance V_i is misspecified. The

variance of $\widehat{\boldsymbol{\beta}}$, provided V_i is properly specified, is equal to

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1}. \tag{4}$$

This variance is the inverse of the Fisher information, and thus represents the amount of information that the data carries about the parameter β. Under the assumption of independence, this variance would be determined as

$$\widetilde{\mathrm{Var}}(\widehat{\beta}) = \left(\sum_{i=1}^{N} X_i' W_i^{-1} X_i\right)^{-1}, \tag{5}$$

with $W_i = \mathrm{diag}(V_i)$ a diagonal matrix with the same values on the diagonal as V_i. The notation $\widetilde{\mathrm{Var}}(\cdot)$ is used to indicate the quantity under independence.

We now define the effective sample size $\widetilde{N}(\beta_k)$ corresponding to a single fixed-effects parameter $\beta_k \in \boldsymbol{\beta}$ as the number of independent measurements that one would need to reach the same amount of information about β_k as in the original data. For the special setting of a model with only an intercept, that is, (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta_0$ and a general homogeneous variance-covariance matrix V_i, we have that the effective sample size is equal to

$$\begin{aligned}\widetilde{N} &= \sum_{i=1}^{N}\left[J_{n_i\times 1}'\left(W_i^{-1/2} V_i W_i^{-1/2}\right)^{-1} J_{n_i\times 1}\right] \\ &= \sum_{i=1}^{N}\left(J_{n_i\times 1}' C_i^{-1} J_{n_i\times 1}\right),\end{aligned} \tag{6}$$

with C_i the correlation matrix and $J_{n_i\times 1}$ an $n_i \times 1$ vector consisting of ones. The derivation of this equation is given in Appendix A.

For a general model, the effective sample size for a parameter β_k can be derived by estimating the weight w to be given to each observational unit such that the variance of β_k in these (weighted) data, under the assumption of independence, that is,

$$\widetilde{\mathrm{Var}}^{w}(\widehat{\beta}_k) = \left[\left(\sum_{i=1}^{N} w X_i' W_i^{-1} X_i\right)^{-1}\right]_{kk},$$

equals the variance in the original data under the assumption of dependence $[\widehat{\mathrm{Var}}(\widehat{\beta}_k)]$. As such, the effective sample size is defined as $\widetilde{N}(\beta_k) = w\sum_i n_i$ with $w = \widehat{\mathrm{Var}}(\widehat{\beta}_k)/\widetilde{\mathrm{Var}}(\widehat{\beta}_k)$ where $\widehat{\mathrm{Var}}(\widehat{\beta}_k)$ and $\widetilde{\mathrm{Var}}(\widehat{\beta}_k)$ are obtained from (4) and (5), respectively.

In the next sections, some special cases of the effective sample size are discussed. First, focus is on models with only an intercept as fixed-effect parameter. Several covariance structures are considered: the compound-symmetry (CS), first-order autoregressive [AR(1)], and three-level hierarchical structures. Second, the specific but important setting of a contrast trend is considered. The various parameters of interest produce important insight.

4.1 Compound-Symmetry Structure

We apply the idea of the effective sample size to the simple but important context of a continuous response Y_{ij} on a set of measurements j which are grouped in a cluster i of size n. Assume the random-intercepts model (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta$. Such a model marginalizes to a so-called *compound-symmetry structure*, where the covariance between two measurements Y_{ij} and Y_{ik} is of the form τ^2, the variance being $\sigma^2 + \tau^2$. The corresponding correlation is $\rho = \tau^2/(\sigma^2 + \tau^2)$, with the variance components as in (3).

The mean parameter β can be estimated as (Laird and Ware 1982):

$$\widehat{\beta} = \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n} Y_{ij}$$

and the variance of $\widehat{\beta}$ equals

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \frac{\sigma^2 + n\tau^2}{Nn}. \tag{7}$$

In the special case that measurements are independent, we would have that $W_i = (\sigma^2 + \tau^2)I_n$ and the variance of $\widehat{\beta}$ would equal

$$\widetilde{\mathrm{Var}}(\widehat{\beta}) = \frac{\sigma^2 + \tau^2}{Nn}. \tag{8}$$

Now, assign a weight w to each observation. The variance under the assumption of independence becomes

$$\widetilde{\mathrm{Var}}^{w}(\widehat{\beta}) = \frac{\sigma^2 + \tau^2}{wNn}. \tag{9}$$

The effective sample size $\widetilde{N}$ can then be calculated by equating (7) and (9):

$$\frac{\sigma^2 + n\tau^2}{Nn} = \frac{\sigma^2 + \tau^2}{wNn},$$

yielding

$$\widetilde{N} = wnN = \frac{nN}{1 + \rho(n-1)}, \tag{10}$$

with $\rho = \tau^2/(\tau^2 + \sigma^2)$. Here, $1 + \rho(n-1)$ is the well-known variance inflation factor in cluster randomized trials. Note that the above equation can also be derived from (6) with $C = \rho J_{n\times n} + (1-\rho)I_n$.

In general, when cluster sizes are not equal, the effective sample size $\widetilde{N}$ for the entire sample equals $\sum_i w_i n_i$, yielding

$$\widetilde{N} = \sum_{i=1}^{N} \frac{n_i}{1 + \rho(n_i - 1)}. \tag{11}$$

In Table 1, the effective sample size $\widetilde{n}(\mathrm{CS})$ for clusters of size $n = 5$ is presented for different correlations ρ. For example, if $\rho = 0.2$, the information obtained from $n = 5$ measurements on the same cluster is similar to what would be obtained from 2.8 independent measurements. There are some interesting special cases. When measurements are independent within a cluster, $\rho = 0$, the effective sample size equals $\widetilde{N} = \sum_i n_i$, the total number of measurements. In case the measurements within a

Table 1. Effective sample size for a cluster of size n with correlation ρ, calculated under the compound-symmetry (CS) model and under the first-order autoregressive [AR(1)] model.

ρ	n	$\widetilde{n}$(CS)	$\widetilde{n}$(AR(1))	ρ	n	$\widetilde{n}$(CS)	$\widetilde{n}$(AR(1))
0	5	5	5	0.5	1	1	1
0.2	5	2.8	3.7	0.5	2	1.33	1.33
0.4	5	1.9	2.7	0.5	5	1.67	2.33
0.6	5	1.5	2.0	0.5	10	1.82	4
0.8	5	1.2	1.4	0.5	100	1.98	34
1	5	1	1	0.5	∞	2	∞

NOTE: Effective sample size calculated as $\widetilde{n}(\text{CS}) = \frac{n}{1+\rho(n-1)}$ and $\widetilde{n}(\text{AR}(1)) = \frac{n-(n-2)\rho}{1+\rho}$.

cluster are perfectly correlated, $\rho = 1$, the effective sample size equals the number of clusters, because $\widetilde{N} = \sum_i \frac{n_i}{n_i} = N$. Further, the right side of Table 1 shows the effective sample size for different cluster sizes (n) and within-correlation $\rho = 0.5$. The effective sample size increases very slowly with growing cluster size, toward the asymptote of $\widetilde{n} = 2$. This will be discussed further in Section 5.

Note that the above derivations are valid for nonnegative correlations. The effective sample size is positive only, and hence well-defined, for correlations $\rho > -1/(n-1)$. Practically, this means that $\rho > -1/(n_i - 1)$ for all trials i. Thus, our argument can be used for mildly negative correlation, down to this bound. Negative correlations are fully acceptable in a marginal interpretation of (3), although values below this bound do not correspond to valid distributions. Notwithstanding this, when a fully hierarchical interpretation is adopted, other than confining attention to the derived marginal model, then negative correlation is not allowable (Verbeke and Molenberghs 2000, 2003).

4.2 Alternative Covariance Structures

We consider the effective sample size for other frequently used correlation structures, when the fixed-effects structure is confined to an intercept. We will consider the independence and first-order stationary autoregressive structures, as well as the one induced by a variance-component specification in a three-level hierarchical model.

When the *independence correlation structure* applies, the effective sample size reduces to $\widetilde{N} = \sum_i n_i$, as would be expected.

A *first-order autoregressive structure* assumes that the covariance between two measurements Y_{ij} and Y_{ik} is of the form $\sigma^2\rho^{|k-j|}$. For a sample with fixed cluster size n, the effective sample size is

$$\widetilde{N} = \frac{n-(n-2)\rho}{1+\rho}N.$$

This can be derived from (6) with $(C)_{kj} = \rho^{|k-j|}$.

In Table 1, the effective sample size for clusters of size $n = 5$ is presented for various correlations ρ, as well as for a cluster of different size n but fixed correlation $\rho = 0.5$. For example, if $\rho = 0.2$, then the information obtained from $n = 5$ measurements on the same individual is similar to what would be obtained from 3.7 independent measurements, which is larger than the effective sample size when a compound-symmetry structure would apply. When $\rho = 0$, the effective sample size reduces to the number of measurements. When $\rho = 1$, the effective sample size is equal to the number of clusters. In the special cases that $n = 1$ or 2 and $\rho = 0$ or 1, the CS and AR(1) structures cannot be distinguished, and hence also the effective sample sizes for both settings are identical. Finally, note that the effective sample size increases faster with growing cluster sizes, as compared with the compound-symmetry structure.

Next, assume a *variance-component structure* in the following three-level model:

$$Y_{ijk} = \beta_0 + u_i + v_{ij} + \varepsilon_{ijk},$$

where β_0 is a fixed-effects parameter, u_i is a random effect at the first level ($i = 1, \ldots, N$), v_{ij} is a random effect at the second level ($j = 1, \ldots, J$), and ε_{ijk} is an error term ($k = 1, \ldots, K$). An example of such a three-level model follows, for example, when measuring paired eyes of an individual in time. The patient is the first level in the data ($i = 1, \ldots, N$), the second level is an eye ($j = 1, 2$) of a patient i, and the third level are the repeated measurements k of eye j of patient i ($k = 1, \ldots, K$). All random terms in the model are assumed to be mutually independent and normally distributed: $u_i \sim N(0, \sigma_u^2)$, $v_{ij} \sim N(0, \sigma_v^2)$, and $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. This model extends (3). In this case, the effective sample size is equal to

$$\widetilde{N} = \sum_{i=1}^{N}\sum_{j=1}^{J} \frac{K}{(m_j - 1)n_i\rho_1 + (n_i - 1)\rho_2 + 1},$$

with n_i the number of eyes measured in patient i ($n_i = 2$), m_j the number of repeated measurements in eye j, ρ_1 and ρ_2 the intraclass correlations between paired eyes and between repeated measurements, respectively. This, in turn, can be derived from (6).

4.3 Contrast Parameter

So far, the effective sample size has been calculated for the overall mean parameter, under a variety of covariance structures. We now switch attention to a contrast parameter, representative of the wide variety of settings where a (treatment) difference is of interest. Assume that individuals are repeatedly measured. Consider (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij}$. The covariate x_{ij} can be either measurement- or individual-specific. In the first case, it changes with i and j; in the second case, it changes with i only.

First, consider the setting where there are n measurements for each individual, together with an individual-specific covariate x_i. The effective sample size for both the intercept β_0 and the parameter β_1 equals (10), the effective sample size for an overall mean parameter, which is in line with intuition. Second, consider the setting of a balanced design, with n measurements for each individual together with measurement-specific covariates $x_{ij} = x_j$. Note that this is a balanced setting with all individuals having the same measurement-specific covariates. The effective sample size for the intercept β_0 then equals

$$\widetilde{N}(\beta_0) = \frac{Nn}{1 + [((n-1)(\sum_j x_j^2) - (\sum_j x_j)^2)/\sum_j x_j^2]\rho},$$

with obvious notation, and for the parameter β_1, $\widetilde{N}(\beta_1) = Nn/(1-\rho)$. The derivation of this formula is given in Appendix B.

As a first example, consider the simple setting where there are two measurements for each individual in a pretest-posttest design. Let x_1 and x_2 denote the covariates at pretest and posttest, respectively. Then $\widetilde{N}(\beta_0)$ reduces to

$$\widetilde{N}(\beta_0) = \frac{2N}{1 + ((x_1 - x_2)^2/(x_1^2 + x_2^2) - 1)\rho}.$$

Simplification for the slope parameter is straightforward: $\widetilde{N}(\beta_1) = 2N/(1-\rho)$, showing that higher positive correlation corresponds to *increasing* information. Indeed, when the correlation is $\rho = 0$, the effective sample size for the contrast parameter is equal to $2N$; $\rho = 0.5$ yields an effective sample size of $4N$; when measurements are perfectly correlated, that is, $\rho = 1$, the effective sample size reaches infinity, meaning that one pair of measurements corresponds to the asymptotic situation of perfect knowledge about the contrast.

As a second example, consider (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij}$ and $x_{ij} = j - 1$. This would correspond to a longitudinal experiment where measurements are taken at regular time points. Then, $\widetilde{N}(\beta_0)$ reduces to

$$\widetilde{N}(\beta_0) = \frac{2Nn(2n-1)}{4n - 2 + \rho(n^2 - 3n + 2)},$$

and $\widetilde{N}(\beta_1) = Nn/(1-\rho)$, as before.

4.4 Binary Data

An important advantage of the effective sample size is that it can be calculated for any parameter and any model family. Let us exemplify this for clustered binary data.

Consider an experiment involving N clusters, the ith of which contains n_i measurements. Suppose Y_{ij} is a binary indicator for the jth measurement of subject i, and $z_i = \sum_{j=1}^{n_i} y_{ij}$ is the total number of positive outcomes in cluster i. Under independence, observations can be seen as binomial counts with

$$\mathrm{E}\left(\frac{z_i}{n_i}\right) = \pi_i, \qquad \mathrm{Var}\left(\frac{z_i}{n_i}\right) = \frac{\pi_i(1-\pi_i)}{n_i}.$$

When responses within a cluster are correlated, this results in extra-binomial variation. It is well known that the variance of z_i/n_i then equals

$$\mathrm{Var}\left(\frac{z_i}{n_i}\right) = \frac{\pi_i(1-\pi_i)}{n_i}[1 + \rho_i(n_i - 1)],$$

with ρ_i the correlation among the observations in cluster i. This correlation can be estimated using a beta-binomial model. As in the setting of a normally distributed response, the effective sample size for $\pi = E[\pi_i]$ takes form (11), as in the normal case. This is because the calculations require up to second moments only, rather than full distributional assumptions.

5. INFORMATION LIMIT

We observed that, for CS, the effective sample size reaches an asymptote when the actual sample size approaches infinity. Thus, there appear to be cases where information for a particular parameter does not grow unboundedly with within-subject sample size. For example, when CS is adopted and focus is on the overall average, an information limit can be derived:

$$\lim_{n\to\infty} \frac{n}{1 + \rho(n-1)} = \frac{1}{\rho}. \tag{12}$$

Thus, under CS, there is a maximum amount of information that can be obtained per subject. Only when observations are independent ($\rho = 0$) is this limit infinity. For example, when $\rho = 0.2$, the limit is equal to 5; hence, for this correlation, a cluster can never contribute more information for the overall mean parameters than would be obtained from five independent measurements. Similarly, when $\rho = 0.5$, a cluster cannot contribute more information than from two independent measurements. This implies that there are no conventional asymptotic arguments possible for $n \to \infty$ in such cases. Of course, in typical longitudinal studies, n_i would be small anyhow, and having $n_i \to \infty$ is less relevant. However, in a meta-analytic context, where i refers to studies, n_i can be quite large, and these considerations become quite relevant.

In contrast to CS, the information limit is infinite when observations follow an AR(1) covariance structure, because

$$\lim_{n\to\infty} \frac{n - (n-2)\rho}{1+\rho} = \infty. \tag{13}$$

Thus, the amount of information, and its behavior as a function of correlation, depends on the covariance structure parameterization, as well as on the parameter values. The contrast between (12) and (13) is dramatic in this respect.

The difference between both situations can intuitively be explained as follows. Under CS, every measurement within a cluster is correlated in the same way with all other measurements. Therefore, there is a limit to what can be learned from a cluster and the additional information coming from new cluster members approaches zero with increasing cluster size. In contrast, under AR(1), the correlation wanes with time lag. So, with time gap between measurements tending to infinity, their correlation tends to zero and hence the additional information tends to that of a new, independent observation. Hence, the infinite information limit.

Also for a contrast parameter, the information limit can be calculated. As an example, consider (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij}$ and $x_{ij} = j$. The information limit for β_0 then equals $4/\rho$. For example, when $\rho = 0.5$, the limit is equal to 8. Already after the sixth measurement, the gain of information about β_0 when adding an additional measurement in a longitudinal experiment is small. In contrast, the information limit about the parameter β_1 is infinite.

6. DEGREES OF FREEDOM IN WALD TESTS

We have considered the concept of effective sample size to compare information, available for a parameter, between correlated and independent situations. Information plays a key role in hypothesis tests, especially when determining denominator degrees of freedom of finite-sample reference distributions. Such distributions, like the t- and F-distributions, are applied when the variance components, present in a test statistic, are unknown and need to be estimated for the data. When data are normally distributed, such reference distributions are exact in the univariate and a variety of dependent cases, such as full multivariate or CS models. This no longer holds for general, often unbalanced, repeated-measures designs and a number of approximate methods are available: Satterthwaite's approximation (Satterthwaite 1941), and the Kenward and Roger (1997) approximation. Note that these methods are only fully developed for the case of linear mixed models and related multivariate normally based models. The accuracy of the approximations depends on the amount of replication. In longitudinal studies, with sufficiently large numbers of patients (large N and small n_i), all approximations typically exhibit similar behavior. However, in a meta-analysis, for example, with merely a few large trials (small N and large n_i), approximations can differ dramatically.

The effective sample size can be used to shed further light on these issues. Our goal for the context of hypothesis testing is twofold. First, by juxtaposing the Satterthwaite and Kenward–Roger methods with an effective sample-size-based version, understanding of all methods' operation and performance will be enhanced. Second, the effective sample-size-based method, unlike the others, can be used for non-Gaussian data as well.

Suppose that inferences are to be made about a single fixed-effects parameter β. The Wald statistic, taking the form

$$T = \frac{\widehat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}}, \tag{14}$$

can be used to test the null hypothesis $H_0 : \beta = 0$. Here, an alternative method to test whether a significant effect exists is proposed, using the effective sample size as building block in the degrees-of-freedom approximation of a scaled Wald test, thus offering important insight regarding the existing methods as well, and in their interconnectedness. Assume that a scaled form $T^* = \lambda T$ of the T-statistic follows a t-distribution with ν degrees of freedom, where λ and ν are unknown quantities. This is similar to the method proposed by Kenward and Roger (1997), where a scaled form of the F-statistic is used for multivariate tests.

Derivation of the scale factor λ follows from matching the first two moments of T^* to the moments of a t-distribution, leading to

$$\lambda^2 = \frac{\nu}{(\nu - 2)V(T)}, \tag{15}$$

with $V(T)$ the variance of the Wald statistic T and $\nu/(\nu - 2)$ the variance of a t-distributed random variable with ν degrees of freedom. The variance $V(T)$ can be approximated by use of the multivariate delta method (see Appendix C). In line with the effective sample size concept, the degrees of freedom are derived from the information arising through the corresponding independent set of data, that is, the sample size minus the number of parameters to be estimated in the fixed-effects structure. This amounts to $\nu = \widetilde{N} - \ell$, with $\widetilde{N}$ the effective sample size. Here, ℓ is the number of parameters.

For the random-intercepts model with a compound-symmetry structure, this leads to

$$\lambda^2 = \frac{\nu}{(\nu - 2)V(T)}, \quad \text{where} \qquad \nu = \sum_{i=1}^{N} \frac{n_i}{1 + (n_i - 1)\rho} - \ell, \tag{16}$$

which is straightforward to compute.

The major difference between the proposed method, and Satterthwaite's or Kenward–Roger's method, is that in the latter methods the degrees of freedom are calculated directly from approximating the distribution for the Wald tests of the individual parameter estimates. The proposed method is more general, in the sense that the concept of the effective sample size is not restricted to a normally distributed response. This is extremely important, because the Satterthwaite and Kenward–Roger methods do not generalize to binary or otherwise non-Gaussian setting. We will return to this in Section 7.3.

Note that the scaled Wald test is defined only when the degrees of freedom are larger than or equal to 2, because the variance of the t-distribution is infinite otherwise. Therefore, in case the calculated degrees of freedom are less than 2, no scaling is applied to the test statistic. This is similar to the Kenward–Roger methodology as implemented in the SAS procedure MIXED. Furthermore, a lower bound of 1 on the degrees of freedom is assumed in the Kenward–Roger methodology.

7. A SIMULATION STUDY

A simulation study was conducted to explore the behavior of the method as proposed in previous sections, and to compare the proposed methodology with (i) the unadjusted test, which uses the t-distribution with the number of measurements minus the number of estimated parameters as the degrees of freedom, (ii) the Satterthwaite method, and (iii) the Kenward–Roger method. Two different normal settings are used. An additional, binary, setting is considered where, of course, no comparison with Satterthwaite or Kenward–Roger is possible. We present the results from each of these in turn. Additional simulations can be found in the supplemental materials.

7.1 The Mean of Compound-Symmetry Correlated Data

In a first simulation study, we generate data from a CS model (3) with $\mathbf{x}'_{ij}\boldsymbol{\beta} = \beta_0$ and an unbalanced design. Three simulation settings are defined assuming that $\beta_0 = 0$, $\sigma^2 = (4, 1, 1)$, and $\tau^2 = (1, 1, 4)$. These settings correspond with an intraclass correlation $\rho = \tau^2/(\sigma^2 + \tau^2)$ of 0.2, 0.5, and 0.8, respectively. We study the t-test for the overall mean corresponding to the null hypothesis $\beta_0 = 0$. For each simulation setting, we vary the mean cluster size and number of clusters.

For each setting, 10,000 sets of data are simulated and for each set the fixed effects are estimated together with the REML

Table 2. Simulation study. Mean of CS correlated data: mean estimated effective sample size ($\widetilde{N}$), mean of scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study corresponding to different settings having N subjects with on average $\overline{n_i}$ measurements. (Unadj: unadjusted; Satterth: Satterthwaite; KR: Kenward–Roger; ess: effective sample size.)

					Observed size			
N	$\overline{n_i}$	ρ	$\widetilde{N}$	λ	Unadj	Satterth	KR	ess
10	4	0.2	19.77	1.02	11.0	5.5	5.5	5.0
		0.5	10.69	1.09	13.0	5.4	5.4	4.3
		0.8	6.07	1.26	11.9	4.7	4.7	3.7
10	2	0.0	18.16	1.05	3.7	3.4	3.4	3.4
		0.2	16.72	1.05	5.3	4.6	4.7	4.6
		0.5	14.02	1.06	6.6	5.0	5.0	5.2
100	4	0.2	230.13	1.00	4.9	4.6	4.6	4.7
		0.5	153.24	1.00	5.0	4.9	4.9	5.0
		0.8	115.92	1.01	5.1	4.9	4.9	4.9
100	2	0.2	147.29	1.00	5.5	5.3	5.3	5.3
		0.5	122.75	1.01	5.8	5.6	5.6	5.6
		0.8	107.39	1.01	5.6	5.4	5.4	5.4
4	100	0.2	42.18	0.96	14.6	5.3	5.3	4.9
		0.5	14.83	1.06	14.8	5.1	5.1	4.4
		0.8	6.97	1.23	14.8	5.0	5.0	4.2
2	100	0.2	56.53	0.99	27.9	11.0	11.0	11.0
		0.5	31.45	1.18	29.5	7.5	7.5	7.5
		0.8	16.93	1.09	29.9	5.7	5.7	5.7
4	10	0.2	19.89	1.02	10.7	5.6	5.6	5.3
		0.5	10.88	1.10	12.9	5.0	4.9	4.2
		0.8	6.21	1.26	12.9	4.7	4.7	3.9
2	10	0.2	12.09	1.20	14.6	10.5	10.5	10.7
		0.5	8.65	1.30	22.2	12.2	13.2	12.2
		0.8	5.80	1.12	26.2	9.3	9.3	11.1

variance estimates of the variance components. Table 2 displays the observed size of a nominal 5% t-test for each of the methods. Additionally, we show the average effective sample size and the average scale factor used in the proposed scaled Wald test.

In a typical longitudinal setting, the number of clusters (individuals) is larger than the number of observations in a cluster (repeated measurements). Simulation results where data are generated under such a setting are presented in the top part of Table 2. The behavior of the proposed method is generally quite good, with an observed size close to the nominal level. The behavior of the ess-based method is comparable to the Satterthwaite and Kenward–Roger methods. The effective sample size decreases with increasing intraclass correlation, as it ought to. The scale parameter is invariably close to 1. When the number of clusters is large, all methods perform equally well, in line with expectation.

In a typical random-effects meta-analytic setting, as in the first example, one encounters a small number of clusters (trials) combined with a large sample size within clusters (number of patients per trial). The lower part of Table 2 presents results for this setting. In most settings, the proposed method works well. However, when both the correlation is large and the number of clusters very small, the ess-based method tends to deteriorate.

Table 3. Simulation study. Mean of AR(1)-correlated data: mean estimated effective sample size ($\widetilde{N}$), mean of scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study, corresponding to different settings having N subjects with n measurements. (Unadj: unadjusted; Satterth: Satterthwaite; KR: Kenward–Roger; ess: effective sample size.)

					Observed size			
N	n	ρ	$\widetilde{N}$	λ	Unadj	Satterth	KR	ess
3	10	0.2	22.98	1.02	7.0	4.7	5.2	4.8
		0.5	13.53	1.05	8.4	4.7	5.9	4.8
		0.8	7.12	1.13	11.7	4.9	6.4	3.9
10	3	0.2	25.20	1.02	6.7	5.2	5.4	5.5
		0.5	17.76	1.04	6.9	5.1	5.5	5.0
		0.8	12.68	1.06	7.1	4.9	5.3	4.1
3	100	0.2	203.02	1.00	6.0	5.6	5.7	5.7
		0.5	103.35	1.00	6.2	5.4	5.8	5.6
		0.8	37.06	1.01	6.7	5.2	5.7	5.2
100	3	0.2	235.07	1.00	6.0	5.8	5.9	5.9
		0.5	167.67	1.00	6.0	5.8	5.9	5.9
		0.8	122.63	1.00	5.9	5.7	5.7	5.6

Note that this is the situation where there is very little information in the data. Also, when we have only two clusters and each cluster contains 10 to 100 observations, the scale parameter can become infinite, especially with large correlation. In addition, the cluster variance will be very poorly estimated, the consequences of which are apparent in Table 2 for $N = 2$. This is due to the infinite variance of a t-distribution when the number of degrees of freedom is smaller than 2, and might occur in situations where there is only a very small amount of information in the data, for example, with an effective sample size smaller than 3.

7.2 The Mean of AR(1)-Correlated Data

Next, consider a longitudinal study with a balanced design and an AR(1) correlation structure. Again, interest is in a test for the overall mean of the response. Various settings for cluster size, number of clusters, and correlation are considered, and results summarized in Table 3.

Also in this setting, the proposed method works very well. Note that the effective sample size under the AR(1) model is much larger when compared to its counterpart under CS, in line with our theoretical developments, underscoring the importance of the correlation structure. Note that the scale parameter λ is virtually 1 when the cluster size is large or when the number of clusters is large.

7.3 Overall Probability of a CS Correlated Binary Response

We now generate binary data from a beta-binomial model, where the outcomes within a cluster follow a binomial distribution, and the cluster-specific success probability is assumed to be drawn from a beta distribution (Molenberghs and Verbeke 2005). In particular, we assume a constant overall mean ($\pi = 0.5$) and constant correlation parameter. Cluster sizes are fixed per simulation setting, and vary between 5 and 50. Given

Table 4. Simulation study. Binary correlated data: mean estimated effective sample size ($\widetilde{N}$), mean of scale parameter (λ), and observed size of nominal 5% Wald t-test from the simulation study for intercept parameter, corresponding to different settings having N subjects with n measurements. (Res: Residual method; ess: effective sample size.)

					Observed size	
N	n	ρ	$\widetilde{N}$	λ	ess	Res
10	5	0.3	25.54	1.02	5.73	7.9
10	5	0.5	18.51	1.03	6.17	9.25
10	5	0.7	14.06	1.04	9.09	12.12
50	5	0.3	116.92	1	5.25	5.73
50	5	0.5	84.87	1.01	4.93	5.45
50	5	0.7	66.43	1.01	4.96	5.48
5	10	0.3	18.82	1.02	8.29	12.97
5	10	0.5	12.39	1.06	8.56	14.52
5	10	0.7	8.84	1.11	5.26	5.26
5	50	0.3	28.05	1	11.00	15.86
5	50	0.5	15.23	1.05	10.88	17.1
5	50	0.7	10.06	1.09	12.86	18.67

that the Satterthwaite and Kenward–Roger methods exist only for normally distributed responses, the residual number of degrees of freedom is often used as an alternative. We will compare this method with our ess-based approach. Results are presented in Table 4. It can be seen that for all settings considered the ess-based method outperforms the residual method. When there is sufficient information in the data, the size of the proposed test is close to 0.05.

8. DATA ANALYSIS

8.1 Cancer of the Ovaries

Consider first a 'naïve' analysis, not accounting for correlation. For the purpose of this illustrative analysis, the small amount of censoring in the data is ignored. Recall that the sample sizes are 411 and 382. The mean log-survival time is estimated to be 0.7906 (s.e. 0.1726). The residual error degrees of freedom, equal to the number of individuals minus the number of parameters to be estimated, equals 792, resulting in $p < 0.0001$ for the t-test corresponding to the null hypotheses of one-year survival. Now, the correlation of individuals within a trial is estimated to be $\widehat{\rho} = 0.038$. Though small, this correlation should be accounted for in this meta-analysis, and has a huge effect on the degrees of freedom. Both Satterthwaite and Kenward–Roger estimated the degrees of freedom equal to 1, resulting in $p = 0.1368$. The effective sample size is estimated as 49. The scaled t-test, with scale parameter $\widehat{\lambda} = 0.30$, has 48 degrees of freedom, resulting in $p = 0.173$.

Further, it should be noted that, due to the correlation, any given trial cannot obtain more information to estimate the mean parameter as corresponds to about 26 independent measurements.

8.2 The Rats Data

In the rats data, the effect of time is estimated as $\widehat{\beta_1} = 0.1934$ (s.e. 0.0059). The residual degrees of freedom are 250. The Satterthwaite and Kenward–Roger methods estimate the degrees of freedom as 214, whereas the effective sample size equals 125.47. All methods result in a significant t-test statistic, with $p < 0.0001$.

9. CONCLUDING REMARKS

We have revisited the effective sample size which is, broadly speaking, the equivalent sample size needed when repeated measures would be uncorrelated, to obtain the same amount of information as in the actual correlated sample.

The use of the ess is threefold. First, it provides a perspective on the amount of information in correlated data. For example, it explains why fixed-effects and random-effects analyses can be so dramatically different in meta-analyses, where the number of trials is typically small, even though the number of patients per trial is commonly large. This is tied to the existence of an information limit for the CS case.

Second, in a hypothesis testing context for a mean in Gaussian models, the ess provides an approximate degrees-of-freedom method that behaves in a very similar way to the Satterthwaite and Kenward–Roger methods. Similarities and differences in performance between the three methods are worth studying in multivariate hypothesis testing contexts as well.

Third, a very practical, and indeed promising use is reserved for testing hypothesis with non-Gaussian, for example, binary, correlated data. Here, the ess can be used, in contrast to Satterthwaite and Kenward–Roger, to derive an approximate degrees-of-freedom method for this context, too.

While the focus in this article has been on the use of the effective sample size in the context of single degree-of-freedom hypothesis testing, it is a much more general concept, worth further exploration. For example, the statistical power calculation and sample size determination for a longitudinal study or study with clustered subjects can be tackled from the perspective of the effective sample size, and is a topic of ongoing research.

APPENDIX A: DERIVATION OF EFFECTIVE SAMPLE SIZE β_0

Assume the special setting of a model with only an intercept, that is, (2) with $\mathbf{x}'_{ij}\boldsymbol{\beta} = \beta_0$ and a homogeneous variance-covariance matrix V_i. The amount of information on the parameter β_0 is represented by the variance

$$\widehat{\mathrm{Var}}(\widehat{\beta_0}) = \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1}, \tag{17}$$

with $X_i = J_{n_i \times 1}$ an $n_i \times 1$ vector of ones. Under the assumption of independence, the variance is

$$\widetilde{\mathrm{Var}}(\widehat{\beta_0}) = \left(\sum_{i=1}^{N} X_i' W_i^{-1} X_i\right)^{-1}. \tag{18}$$

If we assign a weight w_i to each observation, the variance of the weighted dataset under the assumption of independence becomes

$$\widetilde{\mathrm{Var}}^{w}(\widehat{\beta_0}) = \left(\sum_{i=1}^{N} w_i X_i' W_i^{-1} X_i\right)^{-1}. \tag{19}$$

To derive the effective sample size we assume that the variance in the original data under the assumption of dependence is equal to the weighted data under the assumption of independence,

$$\left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} = \left(\sum_{i=1}^{N} w_i X_i' W_i^{-1} X_i\right)^{-1}, \quad (20)$$

or

$$\sum_{i=1}^{N} X_i' V_i^{-1} X_i = \sum_{i=1}^{N} w_i X_i' W_i^{-1} X_i. \quad (21)$$

The weight w_i represents the amount of information a single observation within a cluster contains. The weight w_i corresponding to observations in cluster i is derived from

$$X_i' V_i^{-1} X_i = w_i X_i' W_i^{-1} X_i \quad \text{for all } i = 1, \ldots, N. \quad (22)$$

As a result, $w_i = (X_i' V_i^{-1} X_i)/(X_i' W_i^{-1} X_i)$. The effective sample size is

$$\widetilde{N}(\beta_0) = \sum_{i=1}^{N} w_i n_i = \sum_{i=1}^{N} \frac{J_{1\times n_i} V_i^{-1} J_{n_i\times 1}}{J_{1\times n_i} W_i^{-1} J_{n_i\times 1}} n_i. \quad (23)$$

When V_i is homogeneous, V_i has a constant value on its diagonal, say the value v_i. In this setting, we can write $W_i = v_i I_{n_i}$. As a result, the denominator in (23) is $J_{1\times n_i} W_i^{-1} J_{n_i\times 1} = n_i/v_i$, and $\widetilde{N}(\beta_0)$ is equal to

$$\begin{aligned}\widetilde{N}(\beta_0) &= \sum_{i=1}^{N} v_i \left(J_{1\times n_i} V_i^{-1} J_{n_i\times 1}\right) \\ &= \sum_{i=1}^{N} \left[J_{1\times n_i} \left(W_i^{-1/2} V_i W_i^{-1/2}\right)^{-1} J_{n_i\times 1}\right] \\ &= \sum_{i=1}^{N} \left(J_{1\times n_i} C_i^{-1} J_{n_i\times 1}\right),\end{aligned}$$

with $C_i = W_i^{-1/2} V_i W_i^{-1/2}$.

APPENDIX B: DERIVATION OF EFFECTIVE SAMPLE SIZE β_1

Consider the setting where there are n subject-specific observations for each subject. Interest is in a contrast parameter in the random-intercept model (3) with $\mathbf{x}_{ij}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij}$. Consider the setting of a balanced design, with n measurements for each individual together with measurement-specific covariates $x_{ij} = x_j$.

The variance of the parameters β_0 and β_1 can be derived from

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\beta}) &= \left(\sum_{i=1}^{N} X_i' V_i^{-1} X_i\right)^{-1} \\ &= \left[\sum_i \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \frac{1}{\sigma^2} \times \left(I_n - \frac{\tau^2}{\sigma^2 + n\tau^2} J_{n\times n}\right) \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}\right]^{-1} \\ &= \left[\sum_i \begin{pmatrix} \frac{n}{\sigma^2+n\tau^2} & \frac{1_n x}{\sigma^2+n\tau^2} \\ \frac{1_n x}{\sigma^2+n\tau^2} & \frac{(1_n x^2)(\sigma^2+n\tau^2)-(1_n x)^2\tau^2}{\sigma^2(\sigma^2+n\tau^2)} \end{pmatrix}\right]^{-1} \\ &= \begin{pmatrix} \frac{(1_n x^2)\sigma^2+(n(1_n x^2)-(1_n x)^2)\tau^2}{N(n(1_n x^2)-(1_n x)^2)} & -\frac{(1_n x)\sigma^2}{N(n(1_n x^2)-(1_n x)^2)} \\ -\frac{(1_n x)\sigma^2}{N(n(1_n x^2)-(1_n x)^2)} & \frac{n\sigma^2}{N(n(1_n x^2)-(1_n x)^2)} \end{pmatrix},\end{aligned}$$

with $1_n x = \sum_{j=1}^{n} x_j$ and $1_n x^2 = \sum_{j=1}^{n} x_j^2$. Under the assumption of independence, the variance of the weighted sample equals

$$\begin{aligned}\widetilde{\mathrm{Var}}^{w}(\widehat{\beta}) &= \left[\sum_i \begin{pmatrix} \frac{nw}{\sigma^2+\tau^2} & \frac{w1_n x}{\sigma^2+\tau^2} \\ \frac{w1_n x}{\sigma^2+\tau^2} & \frac{w(1_n x)^2}{\sigma^2+\tau^2} \end{pmatrix}\right]^{-1} \\ &= \begin{pmatrix} \frac{(1_n x^2)(\sigma^2+\tau^2)}{Nw(n(1_n x^2)-(1_n x)^2)} & -\frac{(1_n x)(\sigma^2+\tau^2)}{Nw(n(1_n x^2)-(1_n x)^2)} \\ -\frac{(1_n x)(\sigma^2+\tau^2)}{Nw(n(1_n x^2)-(1_n x)^2)} & \frac{n(\sigma^2+\tau^2)}{Nw(n(1_n x^2)-(1_n x)^2)} \end{pmatrix}.\end{aligned}$$

By equating the variance components corresponding to the intercept, the effective sample size for the intercept is obtained as

$$\begin{aligned}\widetilde{N}(\beta_0) &= \frac{Nn(1_n x^2)(\sigma^2+\tau^2)}{(1_n x^2)\sigma^2 + (n(1_n x^2) - (1_n x)^2)\tau^2} \\ &= \frac{Nn}{1 + (((n-1)(1_n x^2) - (1_n x)^2)/(1_n x^2))\rho}.\end{aligned}$$

Similarly, the effective sample size for the contrast parameter is

$$\widetilde{N}(\beta_1) = Nn\frac{\sigma^2+\tau^2}{\sigma^2} = \frac{Nn}{1-\rho}.$$

APPENDIX C: DERIVATION OF $V(T)$

We are interested in the variance of the test statistic T:

$$V(T) = \widehat{\mathrm{Var}}(\widehat{T}) = \widehat{\mathrm{Var}}\left(\frac{\widehat{\beta}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}}\right).$$

Let us derive this in the context of a compound-symmetry model. The parameters in this model are $(\beta, \sigma^2, \tau^2)$. Using the delta method, the estimated variance of $\widehat{T}$ equals

$$\widehat{\mathrm{Var}}(\widehat{T}) = \begin{pmatrix} \frac{\partial\widehat{T}}{\partial\beta} & \frac{\partial\widehat{T}}{\partial\sigma^2} & \frac{\partial\widehat{T}}{\partial\beta} \end{pmatrix} \widehat{\mathrm{Var}}(\widehat{\beta}, \widehat{\sigma}^2, \widehat{\tau}^2) \begin{pmatrix} \frac{\partial\widehat{T}}{\partial\beta} \\ \frac{\partial\widehat{T}}{\partial\sigma^2} \\ \frac{\partial\widehat{T}}{\partial\beta} \end{pmatrix},$$

where the parameters are replaced by estimates and with derivatives equal to

$$\begin{aligned}\frac{\partial\widehat{T}}{\partial\beta} &= \frac{1}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}}, \\ \frac{\partial\widehat{T}}{\partial\sigma^2} &= -\frac{\widehat{\beta}}{2(\widehat{\mathrm{Var}}(\widehat{\beta}))^{3/2}} \frac{\partial\widehat{\mathrm{Var}}(\widehat{\beta})}{\partial\sigma^2}, \\ \frac{\partial\widehat{T}}{\partial\tau^2} &= -\frac{\widehat{\beta}}{2(\widehat{\mathrm{Var}}(\widehat{\beta}))^{3/2}} \frac{\partial\widehat{\mathrm{Var}}(\widehat{\beta})}{\partial\tau^2}.\end{aligned}$$

Because the variance of $\widehat{\beta}$ in the compound-symmetry model is equal to

$$\widehat{\mathrm{Var}}(\widehat{\beta}) = \left(\sum_{i=1}^{N} \frac{n_i}{\widehat{\sigma}^2 + n_i\widehat{\tau}^2}\right)^{-1},$$

the derivatives of $\widehat{\mathrm{Var}}(\widehat{\beta})$ equal

$$\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \sigma^2} = (\widehat{\mathrm{Var}}(\widehat{\beta}))^2 \left(\sum_{i=1}^{N} \frac{n_i}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right), \tag{24}$$

$$\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \tau^2} = (\widehat{\mathrm{Var}}(\widehat{\beta}))^2 \left(\sum_{i=1}^{N} \frac{n_i^2}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right). \tag{25}$$

Note that, if all sample sizes are equal, that is, $n_i \equiv n$, (24) and (25) reduce to

$$\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \sigma^2} = \frac{1}{Nn},$$

$$\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \tau^2} = \frac{1}{N}.$$

We also obtain

$$\frac{\partial \widehat{T}}{\partial \beta} = \frac{1}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}},$$

$$\frac{\partial \widehat{T}}{\partial \sigma^2} = -\frac{\widehat{\beta}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}}{2} \left(\sum_{i=1}^{N} \frac{n_i}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right),$$

$$\frac{\partial \widehat{T}}{\partial \tau^2} = -\frac{\widehat{\beta}\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta})}}{2} \left(\sum_{i=1}^{N} \frac{n_i^2}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right).$$

Finally, we assume that $\widehat{\beta}$ and $(\widehat{\sigma}^2, \widehat{\tau}^2)$ are uncorrelated, such that the variance of the test statistic is equal to

$$\begin{aligned} V(T) &= 1 + \left(\frac{\partial \widehat{T}}{\partial \sigma^2}\right)^2 \widehat{\mathrm{Var}}(\widehat{\sigma}^2) + \left(\frac{\partial \widehat{T}}{\partial \tau^2}\right)^2 \widehat{\mathrm{Var}}(\widehat{\tau}^2) \\ &\quad + 2\left(\frac{\partial \widehat{T}}{\partial \sigma^2}\right)\left(\frac{\partial \widehat{T}}{\partial \tau^2}\right)\widehat{\mathrm{Cov}}(\widehat{\sigma}^2, \widehat{\tau}^2) \\ &= 1 + \left(\frac{\widehat{\beta}^2 \widehat{\mathrm{Var}}(\widehat{\beta})}{4}\right)\left\{\left(\sum_{i=1}^{N} \frac{n_i}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right)^2 \widehat{\mathrm{Var}}(\widehat{\sigma}^2)\right. \\ &\quad + \left(\sum_{i=1}^{N} \frac{n_i^2}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right)^2 \widehat{\mathrm{Var}}(\widehat{\tau}^2) \\ &\quad + 2\left(\sum_{i=1}^{N} \frac{n_i}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right)\left(\sum_{i=1}^{N} \frac{n_i^2}{(\widehat{\sigma}^2 + n_i\widehat{\tau}^2)^2}\right) \\ &\quad \left. \times \widehat{\mathrm{Cov}}(\widehat{\sigma}^2, \widehat{\tau}^2)\right\}. \end{aligned}$$

In general, we have that

$$V(T) = 1 + \left(\frac{\widehat{\beta}^2}{4\widehat{\mathrm{Var}}(\widehat{\beta})^3}\right)\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\widehat{\beta})],$$

with

$$\begin{aligned} \widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\widehat{\beta})] &= \sum_{l}\left(\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \sigma_l}\right)^2 \widehat{\mathrm{Var}}(\widehat{\sigma}_l) \\ &\quad + \sum_{l}\sum_{k \neq l}\left(\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \sigma_l}\right)\left(\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \sigma_k}\right)\widehat{\mathrm{Cov}}(\widehat{\sigma}_l, \widehat{\sigma}_k) \end{aligned}$$

and

$$\left(\frac{\partial \widehat{\mathrm{Var}}(\widehat{\beta})}{\partial \sigma_l}\right) = \widehat{\mathrm{Var}}(\widehat{\beta})^2 \left(\sum_{i=1}^{N} X_i' \widehat{V}_i^{-1} \frac{\partial \widehat{V}_i}{\partial \sigma_l} \widehat{V}_i^{-1} X_i\right).$$

SUPPLEMENTAL MATERIALS

Data Example, Simulation Results CS Model, and Simulation Intercept and Dose Effect: This file contains an additional illustration of the effective sample size used in small-sample inference, testing for a dose effect in a developmental toxicological experiment, some additional results of the simulation study as described in Section 7.1, showing the behavior of the proposed effective sample size method, and results from a simulation study testing for the dose effect, using the effective sample size methodology. (effective14supplement.pdf; .pdf file)

[Received September 2008. Revised August 2009.]

REFERENCES

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.

Cressie, N. (1991), *Statistics for Spatial Data*, New York: Wiley.

Faes, C., Aerts, M., Geys, H., Molenberghs, G., and Declerck, L. (2004), "Bayesian Testing for Trend in a Power Model for Clustered Binary Data," *Environmental and Ecological Statistics*, 11, 305–322.

Fay, M. P., and Graubard, B. I. (2001), "Small-Sample Adjustments for Wald-Type Tests Using Sandwich Estimators," *Biometrics*, 57, 1198–1206.

Giesbrecht, G. F., and Burns, J. C. (1985), "Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results," *Biometrics*, 41, 853–862.

Johnson, R. A., and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, Upper Saddle River, NJ: Prentice Hall.

Jones, B., and Kenward, M. G. (2003), *The Analysis of Cross-Over Studies* (2nd ed.), London: Chapman & Hall.

Kenward, M. G., and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects From Restricted Maximum Likelihood," *Biometrics*, 53, 983–997.

Laird, N. M., and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Molenberghs, G., and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, New York: Springer.

Ovarian Cancer Meta-Analysis Project (1991), "Cyclophosphamide Plus Cisplatin versus Cyclophosphamide, Doxorubicin, and Cisplatin Chemotherapy of Ovarian Carcinoma: A Meta-Analysis," *Journal of Clinical Oncology*, 9, 1668–1674.

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985), "The Developmental Toxicity of Ethylene Glycol in Rats and Mice," *Toxicology and Applied Pharmacology*, 81, 113–127.

Satterthwaite, F. E. (1941), "Synthesis of Variance," *Psychometrika*, 6, 309–316.

Skinner, C. J., Holt, D., and Smith, T. M. F. (1989), *Analysis of Complex Surveys*, New York: Wiley.

Thiebaux, H. J., and Zwiers, F. W. (1984), "The Interpretation and Estimation of Effective Sample Size," *Journal of Climate and Applied Meteorology*, 23, 800–811.

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer.

——— (2003), "The Use of Score Tests for Inference on Variance Components," *Biometrics*, 59, 254–262.

Verdonck, A., De Ridder, L., Verbeke, G., Bourguignon, J. P., Carels, C., Kuhn, E. R., Darras, V., and de Zegher, F. (1998), "Comparative Effects of Neonatal and Prepubertal Castration on Craniofacial Growth in Rats," *Archives of Oral Biology*, 43, 861–871.

Welsh, A. H. (1996), *Aspects of Statistical Inference*, New York: Wiley.