Proc. Estonian Acad. Sci. Phys. Math., 2005, **54**, 4, 243–254

A new approximation to the variance of the ANOVA estimate of the intraclass correlation coefficient

Tanel Kaart

Institute of Animal Science, Estonian Agricultural University, Kreutzwaldi 1, 51014 Tartu, Estonia; ktanel@eau.ee

Received 24 November 2004, in revised form 5 April 2005

Abstract. The simple random one-way linear model and ANOVA estimators of the intraclass correlation coefficient are examined. New approximation for the sampling variance of the intraclass correlation coefficient is derived and its minimum, corresponding to a balanced data set, is established. The theoretical results are checked with simulation experiments. In addition, the effect of data set imbalance and structure on the accuracy of intraclass correlation coefficient estimators is studied by modelling.

Key words: variance analysis, intraclass correlation coefficient, heritability, sampling variance, data imbalance, optimal design.

1. INTRODUCTION

The parameters calculated in genetical applications of mixed linear models are usually different ratios of variance components, showing the proportions of variability of observations caused by specified factors. In general, these ratios are called intraclass correlation coefficients. Different linear combinations of these coefficients are used in applications. For example, in genetic studies based on half-sib families, the intraclass correlation coefficient measures one quarter of the additive genetic contribution, called the heritability coefficient of the observed trait. Several heritability coefficients used in population genetics are discussed in [1,2], for example.

In spite of a large number of applications, the estimation theory of the intraclass correlation coefficient has many unsolved problems. Only few articles are available

about the accuracy of estimates and about the effect of data structure. Moreover, all these papers focus on balanced data designs only (see [3,4]). In the present paper we estimate the accuracy of the estimators of the intraclass correlation coefficient in unbalanced designs but assuming the simplest mixed linear model – the one-way random model. We also estimate the effect of data imbalance and structure on the accuracy of the intraclass correlation coefficient. Theoretical results are illustrated with statistical modelling.

2. THE MODEL AND THE ESTIMATES

In the following we use the matrix notation in terms of matrix blocks accepted, e.g., in [5,6]. For instance, $\mathbf{A} = \{_{\mathrm{d}}\mathbf{A}_i\}_{i=1}^{t}$ is an $r \times s$ block-diagonal matrix with $r_i \times s_i$ matrices $\mathbf{A}_i$ on the main diagonal, $r = r_1 + \ldots + r_t$ and $s = s_1 + \ldots + s_t$.

Consider the mixed linear model

$$y_{ij} = \mu + u_i + e_{ij} \tag{1}$$

or, in matrix notation,

$$\mathbf{y} = \mathbf{1}_N\mu + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$ is the $N \times 1$ vector of observed values, μ is the only fixed effect in the model (the mean), $\mathbf{1}'_N = (1 \ldots 1)'_N$ and $\mathbf{Z} = \{_{\mathrm{d}}\mathbf{1}_{n_i}\}_{i=1}^{a}$ are known design matrices of order $N \times 1$ and $N \times a$, respectively, associating fixed and random effects with $\mathbf{y}$, $\mathbf{u}' = (u_1 \ldots u_a)'$ is a vector of random effects, $\mathbf{e}$ is an $N \times 1$ vector of random residuals. The number of levels in a random factor is traditionally marked as a, and the number of objects per the ith level in the one-way model is denoted by n_i.

The expectation and the variance-covariance structure are represented as

$$\mathrm{E}(\mathbf{y}) = \mu, \quad \mathrm{Var}(\mathbf{u}) = \sigma_u^2\mathbf{I}_a, \quad \mathrm{Var}(\mathbf{e}) = \sigma_e^2\mathbf{I}_N, \quad \mathrm{Cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$$

and

$$\mathbf{V} = \mathrm{Var}(\mathbf{y}) = \{_{\mathrm{d}}\sigma_u^2\mathbf{J}_{n_i} + \sigma_e^2\mathbf{I}_{n_i}\}_{i=1}^{a}, \tag{2}$$

where $\mathbf{I}$ and $\mathbf{J}$ denote the identity matrix and square matrix of ones, respectively.

It is assumed that the effects u_i and e_{ij} in the model (1) are independently and normally distributed so that

$$u_i \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2), \quad (i = 1, 2, ..., a;\ j = 1, 2, ..., n_i). \tag{3}$$

It is well known that the sum of squares corresponding to the main effect and expressed as

$$\mathrm{SS}(u) = \mathbf{y}' \left[\{_{\mathrm{d}}\mathbf{J}_{n_i}/n_i\}_{i=1}^{a} - \mathbf{J}_N/N \right] \mathbf{y} = \mathbf{y}'\mathbf{Q}_1\mathbf{y}, \tag{4}$$

and the sum of squares corresponding to the error term and expressed as

$$\mathrm{SS}(e) = \mathbf{y}' \left[\mathbf{I}_N - \{{}_{\mathrm{d}}\mathbf{J}_{n_i}/n_i\}_{i=1}^{a}\right] \mathbf{y} = \mathbf{y}'\mathbf{Q}_2\mathbf{y},$$

are statistically independent and that $\mathrm{SS}(e)/\sigma_e^2$ is Chi-squared distributed:

$$\mathrm{SS}(e)/\sigma_e^2 \sim \chi^2_{N-a}. \tag{5}$$

If the data set is balanced, that is, $n_i = n$ for $i = 1, \ldots, a$, then it also holds that

$$\mathrm{SS}(u)/(n\sigma_u^2 + \sigma_e^2) \sim \chi^2_{a-1}.$$

In the case of unbalanced data the last distribution is not true. However, as noted in [7], the formulas derived in [8] can be used to express the quadratic form (4) as a linear combination of independent central Chi-squared variables of the form

$$\mathrm{SS}(u) = \mathbf{y}'\mathbf{Q}_1\mathbf{y} \sim \sum_{i=1}^{s} \lambda_i \chi^2_{m_i}, \tag{6}$$

where $\lambda_1, \lambda_2, ..., \lambda_s$ are the distinct nonzero eigenvalues of $\mathbf{Q}_1\mathbf{V}$ with multiplicities $m_1, m_2, ..., m_s$, respectively, and $\mathbf{V}$ is the variance matrix of observed values defined by Eq. (2). As the further operations with the mixture distribution (6) are complicated, an approximation, based on Satterthwaite's procedure [9] and presented in [7], is used in the form

$$\sum_{i=1}^{s} \lambda_i \chi^2_{m_i} \approx \lambda \chi^2_m, \tag{7}$$

where

$$\lambda = \frac{\sum_{i=1}^{s} m_i \lambda_i^2}{\sum_{i=1}^{s} m_i \lambda_i}, \qquad m = \frac{\left(\sum_{i=1}^{s} m_i \lambda_i\right)^2}{\sum_{i=1}^{s} m_i \lambda_i^2}. \tag{8}$$

The approximation is exact when the data set is balanced, that is, when $n_i = n$ for $i = 1, \ldots, a$ [7].

The ANOVA estimators of variance components σ_u^2 and σ_e^2 are obtained by equating the mean squares with their expected values and are expressed as

$$\hat{\sigma}_u^2 = \frac{1}{d}\left[\mathrm{MS}(u) - \mathrm{MS}(e)\right]$$

and

$$\hat{\sigma}_e^2 = \mathrm{MS}(e),$$

where $\mathrm{MS}(u) = \mathrm{SS}(u)/(a-1)$, $\mathrm{MS}(e) = \mathrm{SS}(e)/(N-a)$, and

$$d = \frac{1}{a-1}\left(N - \frac{1}{N}\sum_{i=1}^{a} n_i^2\right). \tag{9}$$

The estimator $\hat{\rho}$ of the intraclass correlation coefficient ρ, which measures the magnitude of random effects, is calculated as the ratio of variances:

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} = \frac{\mathrm{MS}(u) - \mathrm{MS}(e)}{\mathrm{MS}(u) + (d-1)\mathrm{MS}(e)}. \tag{10}$$

3. THE ACCURACY OF THE ESTIMATED INTRACLASS CORRELATION COEFFICIENT

There is no exact formula for the variance of the intraclass correlation coefficient estimate even in the balanced case. Usually an approximate formula is used:

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2[1+(n-1)\rho]^2(1-\rho)^2}{n(n-1)(a-1)},$$

derived in [10] using a first-order Taylor-series expansion of the equality (10) with n replacing d. In unbalanced data the approximation of $\mathrm{Var}(\hat{\rho})$ was published in [11] and has the following form:

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2(N-1)(1-\rho)^2[1+(d-1)\rho]^2}{d^2(N-a)(a-1)}. \tag{11}$$

Derivation of this formula is based on an approximate formula for the variance of the ratio of two random variables:

$$\mathrm{Var}(y/x)$$
$$\approx [\mathrm{E}(y)/\mathrm{E}(x)]^2 \left\{\mathrm{Var}(y)/[\mathrm{E}(y)]^2 + \mathrm{Var}(x)/[\mathrm{E}(x)]^2 - 2\mathrm{Cov}(y, x/[\mathrm{E}(y)\mathrm{E}(x)])\right\},$$

applied to the estimate of the intraclass correlation coefficient expressed through sums of squares.

Next an alternative expression for $\mathrm{Var}(\hat{\rho})$ is derived based on approximations (7) and on the first-order Taylor series expansion of the variance of the nonlinear function of parameter w estimator of the form

$$\mathrm{Var}[f(\hat{w})] \approx [\partial f(w)/\partial w]^2 \mathrm{Var}(\hat{w}), \tag{12}$$

where the derivative is evaluated at the mean of $\hat{w}$.

Theorem 1. *In a one-way random model under the normality assumptions* (3)*, the variance of the intraclass correlation coefficient estimate can approximately be expressed as*

$$\mathrm{Var}(\hat{\rho}) \approx \frac{2m\lambda^2(N-a)^2(N-a+m-2)(1-\rho)^4}{d^2(a-1)^2(N-a-2)^2(N-a-4)\sigma_e^4}, \tag{13}$$

where m, λ, and d are defined by formulas (8) *and* (9)*.*

Proof. Let $\hat{w} = \mathrm{MS}(u)/\mathrm{MS}(e)$. Then, on the basis of (10),

$$\hat{\rho} = \frac{\hat{w} - 1}{\hat{w} + d - 1} = f(\hat{w}). \tag{14}$$

Following (5)–(7), we get that $\hat{w}$ is approximately distributed as follows:

$$\hat{w} \sim \frac{m\lambda}{(a-1)\sigma_e^2} F_{m,N-a},$$

where $F_{m,N-a}$ is the F-distribution having m and $N-a$ degrees of freedom. Here, depending on the context, F denote both the distribution and random variable with the F-distribution. Because of

$$\mathrm{Var}(F_{m,N-a}) = 2(N-a)^2(N-a+m-2)\big/[m(N-a-2)^2(N-a-4)]$$

we have

$$\mathrm{Var}(\hat{w}) \approx \frac{(m\lambda)^2}{(a-1)^2\sigma_e^4} \, \frac{2(N-a)^2(N-a+m-2)}{m(N-a-2)^2(N-a-4)}.$$

From (14) it follows that

$$\frac{\partial f(\hat{w})}{\partial \hat{w}} = \frac{d}{(\hat{w} + d - 1)^2},$$

or, because $\hat{w} = [1 + (d-1)\hat{\rho}]/(1-\hat{\rho})$,

$$\frac{\partial f(\hat{w})}{\partial \hat{w}} = \frac{(1-\hat{\rho})^2}{d},$$

and, based on the approximation (12),

$$\begin{aligned} \mathrm{Var}(\hat{\rho}) &\approx \frac{(1-\rho)^4}{d^2} \, \frac{(m\lambda)^2}{(a-1)^2\sigma_e^4} \, \frac{2(N-a)^2(N-a+m-2)}{m(N-a-2)^2(N-a-4)} \\ &= \frac{2m\lambda^2(N-a)^2(N-a+m-2)(1-\rho)^4}{d^2(a-1)^2(N-a-2)^2(N-a-4)\sigma_e^4}, \end{aligned}$$

which completes the proof of Theorem 1. □

The real data analysis uses mainly the square root of the sampling variance (sampling standard deviation) of a parameter estimate. The reason is that standard deviations are easier to interpret and they are also the basis for the accuracy and significance testing of the estimation procedures.

4. A SIMULATION STUDY

We carried out a simulation study to investigate the accuracy of the derived formula, applied in estimating the standard deviation $\sigma(\hat{\rho})$ of the intraclass correlation coefficient estimate $\hat{\rho}$. The data size $N = 360$ was used as a typical number in small practical experiments. A practical advantage of this sample size is that it can be divided in different ways into smaller groups of equal size. The number of groups, a, used in simulations, was taken 4, 20, and 90. Random effects were generated by normal distributions (3). Without the loss of generality, a value of $\sigma_e^2 = 1$ was used throughout the simulations. In generation of random effects u_i, the intraclass correlation coefficient ρ was taken equal to 0.0125, 0.0625, 0.15, and 0.8. Small values of ρ were chosen because only these have a real meaning in most applications. One larger value ($\rho = 0.8$) was chosen to verify the accuracy of the derived formula near the upper limit of intraclass correlation coefficient values. To control the expressions of $\sigma(\hat{\rho})$ in the case of different data designs $\mathbf{D} = \{n_1, n_2, \ldots, n_a\}$, the measure of design imbalance introduced in [12] in the form

$$\nu(\mathbf{D}) = 1 \Big/ a \sum_{i=1}^{a} \left(\frac{n_i}{N}\right)^2$$

was used. Hence, $1/a < \nu(\mathbf{D}) \leq 1$, and the measure $\nu(\mathbf{D})$ attains its maximum value 1 if and only if the design **D** is balanced. The algorithm for generating designs with a specified degree of imbalance proposed in [7], pp. 76–80, was realized in SAS Interactive Matrix Language [13] and used to generate data sets with three specified degrees of imbalance: $\nu = 0.3$, 0.6, 0.9. Due to the very intensive computer calculations, only 1000 simulations were made with each combination of the values of the parameters a, ρ, and ν. This modelling size provided an idea of general tendencies. The parameters compared were (a) the observed standard deviation $\sigma(\hat{\rho})$ of the estimated intraclass correlation coefficient; (b) the predicted standard deviation $\sigma_S(\hat{\rho}|\rho)$ of the intraclass correlation coefficient estimate calculated as the square root of the approximation (11); (c) the predicted standard deviation $\sigma_K(\hat{\rho}|\rho)$ of the intraclass correlation coefficient estimate calculated as the square root of the approximation (13); (d) the estimated standard deviation $\widehat{\sigma}_S(\hat{\rho}|\hat{\rho})$ of the intraclass correlation coefficient estimate calculated as the square root of the approximation (11) with $\hat{\rho}$ substituted for ρ; (e) the estimated standard deviation $\widehat{\sigma}_K(\hat{\rho}|\hat{\rho})$ of the intraclass correlation coefficient estimate calculated as the square root of the approximation (13) with $\hat{\rho}$ substituted for ρ. The simulation results are presented in Table 1.

The results in Table 1 show that in the case of small values of ρ, both the approximations (11) and (13) give quite similar results that seem to be unbiased. For large intraclass correlation coefficients the formula (10) underestimates the real values of ρ. The estimate $\widehat{\sigma}_S(\hat{\rho}|\hat{\rho})$ based on the expression (11), and the estimate $\widehat{\sigma}_K(\hat{\rho}|\hat{\rho})$ based on the expression (13), underestimate $\sigma(\hat{\rho})$ when the number of

groups is small and the intraclass correlation coefficient is large. The difference from simulated values is smaller by using the expression (13).

Table 1. The observed, predicted, and estimated standard deviations of the intraclass correlation coefficient estimate $\hat{\rho}$ for different ρ values, data set imbalances ν, and numbers of groups a. The data size $N = 360$ and residual variance $\sigma_e^2 = 1$ were kept constant.

ρ	ν	a	$\mathrm{E}(\hat{\rho})$	$\sigma(\hat{\rho})^*$	$\sigma_S(\hat{\rho}\|\rho)^\dagger$	$\widehat{\sigma}_S(\hat{\rho}\|\hat{\rho})^\ddagger$	$\sigma_K(\hat{\rho}\|\rho)^\dagger$	$\widehat{\sigma}_K(\hat{\rho}\|\hat{\rho})^\ddagger$
0.0125	0.3	4	0.0147	0.0468	0.0460	0.0459	0.0471	0.0531
		20	0.0122	0.0249	0.0257	0.0261	0.0254	0.0264
		90	0.0133	0.0459	0.0517	0.0459	0.0460	0.0466
	0.6	4	0.0118	0.0209	0.0210	0.0201	0.0221	0.0221
		20	0.0129	0.0225	0.0286	0.0240	0.0234	0.0239
		90	0.0140	0.0471	0.0446	0.0444	0.0451	0.0453
	0.9	4	0.0116	0.0183	0.0193	0.0183	0.0196	0.0188
		20	0.0118	0.0214	0.0371	0.0232	0.0225	0.0222
		90	0.0139	0.0449	0.0485	0.0529	0.0449	0.0448
0.0625	0.3	4	0.0587	0.0794	0.0806	0.0730	0.0897	0.0849
		20	0.0608	0.0418	0.0381	0.0371	0.0451	0.0442
		90	0.0631	0.0496	0.0636	0.0597	0.0511	0.0514
	0.6	4	0.0585	0.0592	0.0580	0.0523	0.0653	0.0591
		20	0.0620	0.0378	0.0616	0.0627	0.0397	0.0389
		90	0.0602	0.0491	0.0579	0.0600	0.0494	0.0491
	0.9	4	0.0620	0.0543	0.0563	0.0536	0.0582	0.0554
		20	0.0641	0.0361	0.0780	0.0739	0.0373	0.0371
		90	0.0607	0.0479	0.0906	0.0855	0.0488	0.0484
0.15	0.3	4	0.1231	0.1210	0.1312	0.1054	0.1563	0.1250
		20	0.1453	0.0706	0.0628	0.0598	0.0777	0.0739
		90	0.1448	0.0585	0.0740	0.0715	0.0606	0.0597
	0.6	4	0.1377	0.1097	0.1126	0.0959	0.1309	0.1113
		20	0.1477	0.0613	0.0777	0.0750	0.0639	0.0620
		90	0.1462	0.0549	0.0825	0.0805	0.0561	0.0555
	0.9	4	0.1364	0.0979	0.1114	0.0959	0.1156	0.0994
		20	0.1472	0.0594	0.0882	0.0841	0.0639	0.0619
		90	0.1490	0.0528	0.0816	0.0821	0.0543	0.0539
0.8	0.3	4	0.6614	0.2230	0.1407	0.1555	0.1691	0.1879
		20	0.7728	0.0821	0.0765	0.1010	0.0844	0.0894
		90	0.7950	0.0406	0.0598	0.0589	0.0442	0.0446
	0.6	4	0.6903	0.2034	0.1395	0.1496	0.1554	0.1685
		20	0.7770	0.0640	0.1722	0.1703	0.0657	0.0697
		90	0.7945	0.0332	0.0572	0.0612	0.0350	0.0355
	0.9	4	0.7042	0.1922	0.1433	0.1534	0.1370	0.1476
		20	0.7823	0.0640	0.4972	0.5041	0.0944	0.0991
		90	0.7965	0.0309	0.0294	0.0297	0.0308	0.0311

* Observed standard deviations $\sigma(\hat{\rho})$ were found from 1000 replicated samples.

† Predicted standard deviations $\sigma_S(\hat{\rho}|\rho)$ and $\sigma_K(\hat{\rho}|\rho)$ were calculated as square roots of formulas (11) and (13), respectively.

‡ Estimated standard deviations $\widehat{\sigma}_S(\hat{\rho}|\hat{\rho})$ and $\widehat{\sigma}_K(\hat{\rho}|\hat{\rho})$ were calculated as square roots of formulas (11) and (13), respectively, with $\hat{\rho}$ substituted for ρ.

5. THE EFFECT OF DATA IMBALANCE

In the following it is proved that $\mathrm{Var}(\hat{\rho})$ expressed by (13) attains its minimum if and only if the data set is balanced.

Theorem 2. *For fixed values of N, a, σ_u^2, and σ_e^2, $\mathrm{Var}(\hat{\rho})$ attains a minimum if and only if the data set is balanced.*

Proof. Rewrite the expression (13) in the form

$$\mathrm{Var}(\hat{\rho}) \approx \left[\frac{m\lambda^2(N-a-2)}{d^2(a-1)^2} + \frac{(m\lambda)^2}{d^2(a-1)^2}\right] \frac{2(N-a)^2(1-\rho)^4}{(N-a-2)^2(N-a-4)\sigma_e^4}, \tag{15}$$

where only the first part depends on the design. From formulas (8) and general properties of eigenvalues we have

$$m\lambda = \sum_{i=1}^{s} m_i\lambda_i = \mathrm{tr}(\mathbf{Q}_1\mathbf{V}), \qquad m\lambda^2 = \sum_{i=1}^{s} m_i\lambda_i^2 = \mathrm{tr}\left[(\mathbf{Q}_1\mathbf{V})^2\right]. \tag{16}$$

From formula (4) we have

$$\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) = \left(N - \frac{1}{N}\sum_{i=1}^{a} n_i^2\right)\sigma_u^2 + (a-1)\sigma_e^2.$$

As we also have the expression for d of the form (9), we can write

$$\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) = (a-1)(\sigma_e^2 + d\sigma_u^2),$$

from which it follows that

$$d = \frac{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})}{\sigma_u^2(a-1)} - \frac{\sigma_e^2}{\sigma_u^2} = \frac{1}{\sigma_u^2(a-1)}\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) - (a-1)\sigma_e^2\right]. \tag{17}$$

From the expressions (16) and (17) we have for the first addend in the square brackets of (15) that

$$\begin{aligned}\frac{m\lambda^2(N-a-2)}{d^2(a-1)^2} &= \frac{\sigma_u^4(N-a-2)\mathrm{tr}\left[(\mathbf{Q}_1\mathbf{V})^2\right]}{\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V}) - (a-1)\sigma_e^2\right]^2} \\ &= \sigma_u^4(N-a-2)\frac{\mathrm{tr}\left[(\mathbf{Q}_1\mathbf{V})^2\right]}{\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V})\right]^2}\left[1 - \frac{(a-1)\sigma_e^2}{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})}\right]^{-2}.\end{aligned}$$

Here the first term does not depend on design. The second term has its minimum value equal to the reciprocal of the rank of $\mathbf{Q}_1$, which is equal to $a-1$, if and only if $n_i = n$ for all i ([14], p. 303). For the third term we have

$$\left[1 - \frac{(a-1)\sigma_e^2}{\mathrm{tr}(\mathbf{Q}_1\mathbf{V})}\right]^{-2} = \left[1 - \frac{\sigma_e^2}{\sigma_e^2 + d\sigma_u^2}\right]^{-2} = \left(1 + \frac{\sigma_e^2}{d\sigma_u^2}\right)^2,$$

which has its minimum, equal to $(1+\sigma_e^2/n\sigma_u^2)^2$, if and only if $n_i = n$ for all i, because d is at its maximum, namely $d = n$, if and only if $n_i = n$ for all i.

Similarly, we have for the second addend in the square brackets of (15) that

$$\frac{(m\lambda)^2}{d^2(a-1)^2} = \frac{\sigma_u^4\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V})\right]^2}{\left[\mathrm{tr}(\mathbf{Q}_1\mathbf{V})\right]^2}\left(1+\frac{\sigma_e^2}{d\sigma_u^2}\right)^2 = \sigma_u^4\left(1+\frac{\sigma_e^2}{d\sigma_u^2}\right)^2.$$

The last expression here has its minimum $\sigma_u^4(1+\sigma_e^2/n\sigma_u^2)^2$ if and only if $n_i = n$ for all i. We therefore conclude that $\mathrm{Var}(\hat{\rho})$ is at its minimum, which is given by the formula

$$\begin{aligned}
\mathrm{Var}(\hat{\rho}) &\approx \left[\frac{\sigma_u^4(N-a-2)}{(a-1)}\left(1+\frac{\sigma_e^2}{n\sigma_u^2}\right)^2 + \sigma_u^4\left(1+\frac{\sigma_e^2}{n\sigma_u^2}\right)^2\right] \\
&\quad \times \frac{2(N-a)^2(1-\rho)^4}{(N-a-2)^2(N-a-4)\sigma_e^4} \\
&= \frac{2(\sigma_e^2+n\sigma_u^2)^2(1-\rho)^4(N-a)^2(N-3)}{\sigma_e^4 n^2(a-1)(N-a-2)^2(N-a-4)} \\
&= \frac{2[1+(n-1)\rho]^2(1-\rho)^2(N-a)^2(N-3)}{n^2(a-1)(N-a-2)^2(N-a-4)}, \qquad (18)
\end{aligned}$$

if and only if the data set is balanced. This completes the proof of Theorem 2. □

Note that the minimum of $\mathrm{Var}(\hat{\rho})$ expressed by (18) and corresponding to balanced data has the same form as approximated $\mathrm{Var}(\hat{\rho})$ derived in [15] assuming balanced data.

To visualize the effect of data imbalance on the accuracy of the estimated intraclass correlation coefficient, modelling experiments were used. Standard deviations of the intraclass correlation coefficient estimate were calculated by formula (13) in the case of different combinations of data imbalance and intraclass correlation coefficients. Since there are different data designs corresponding to a given imbalance, on an average five designs were generated with each specified $\nu(\mathbf{D})$ value, and the average value of $\sigma(\hat{\rho})$ was used to characterize the effect of the corresponding imbalance. Figure 1 shows the dependence of the standard deviation of the estimated intraclass correlation coefficient on $\nu(\mathbf{D})$ and on ρ, keeping the data size $N = 360$, number of groups $a = 20$, and error variance $\sigma_e^2 = 1$.

The modelling results show that even a quite notable increase in data imbalance does practically not reduce the accuracy of the intraclass correlation coefficient estimate. Yet, in case of very imbalanced data, the standard deviation of $\hat{\rho}$ increases quickly with imbalance. In modelling experiments with other numbers of groups (not shown here) the influence of data imbalance on the accuracy of intraclass correlation coefficient estimators was stronger at a small number of groups. It is also observed that in case of very unbalanced design the accuracy of $\hat{\rho}$ decreases

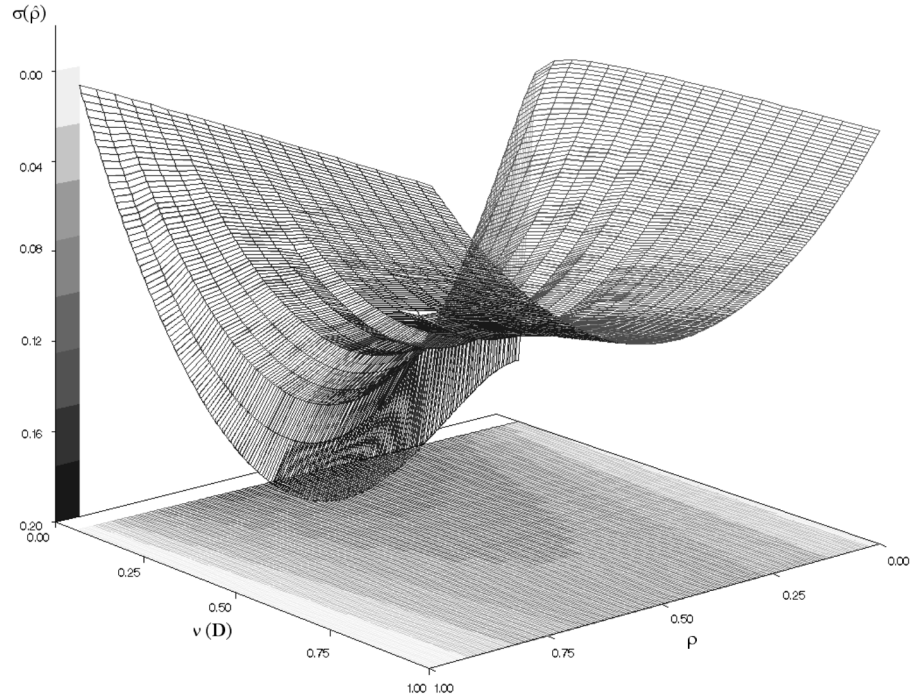

Fig. 1. The pattern of $\sigma(\hat{\rho})$ in different data set imbalances and true intraclass correlation coefficient values for $N = 360$, $a = 20$, $\sigma_e^2 = 1$. Note that the y-axis is directed downward.

more drastically at small values of the intraclass correlation coefficient. The most imprecise estimates were obtained for average values of ρ.

6. OPTIMAL DESIGN

The number of objects per group, which minimizes the sampling variance of σ_u^2, is derived in [16] for balanced designs, considering the group size n as a continuous argument and studying the derivatives of the expression of $\mathrm{Var}(\hat{\sigma}_u^2)$. The minimum $\mathrm{Var}(\hat{\sigma}_u^2)$, and also $\sigma(\hat{\sigma}_u^2)$, is obtained by considering

$$n = \frac{N(\tau + 1) + 1}{N\tau + 2}$$

observations per group, where $\tau = \sigma_u^2 / \sigma_e^2$. Modelling experiments with different data sizes, numbers of groups, and values of the intraclass correlation coefficient showed that the same group sizes guarantee the smallest values of $\mathrm{Var}(\hat{\rho})$ and $\sigma(\hat{\rho})$.

Figure 2 shows the pattern of $\sigma(\hat{\rho})$, calculated as the square root of the expression (18), and optimum number of observations per group, found for a fixed

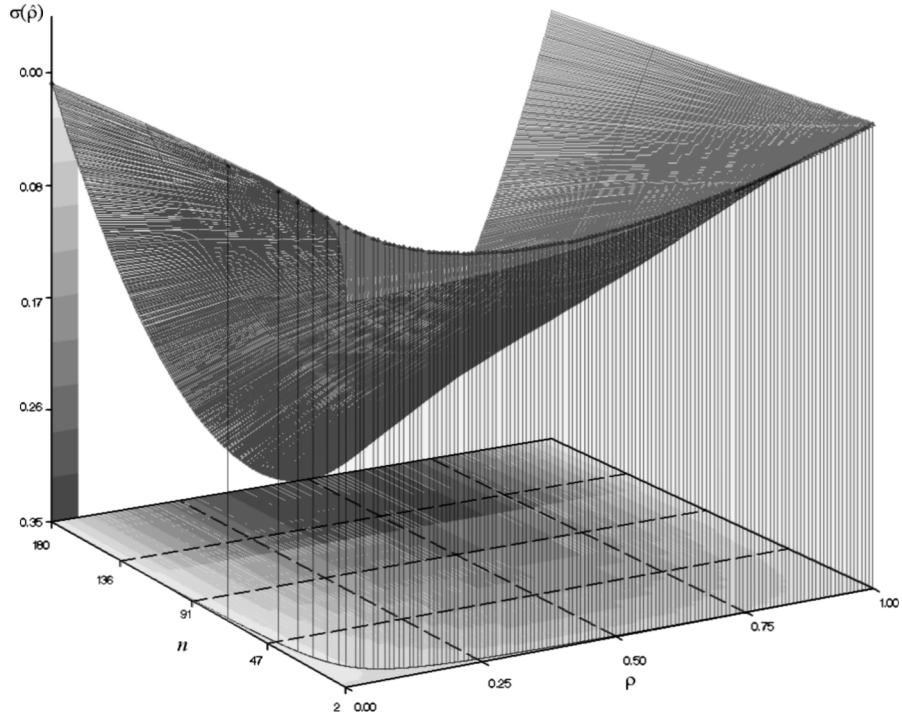

Fig. 2. The pattern of $\sigma(\hat{\rho})$ and the optimal number of observations per group (vertical arrows for integer numbers and dotted line on the xy-plane for continuous numbers) in different true intraclass correlation coefficient values ($N = 360$, $\sigma_e^2 = 1$). Note that the y-axis is directed downward.

data size $N = 360$ in the case of different group sizes and intraclass correlation coefficient values. Based on Fig. 2 and on the results presented in Table 1, the following conclusions were made: (1) the effect of data design is the smallest, when the values of the intraclass correlation coefficient are close to its theoretical limits; (2) a small number of groups, even with a large number of observations, may cause dramatic loss of accuracy, except in the case of small values of ρ. In case of the latter (usual in, for example, genetic studies) a very small number of observations per group should be avoided.

ACKNOWLEDGEMENTS

This investigation was partially performed during the visit to the Linnaeus Centre for Bioinformatics, supported by the European Commission programme Human Research Potential & the Socio-economic Knowledge Base: Access to Research Infrastructures, project number HPRI-CT-2001-00153.

REFERENCES

1. Shen, P.-S., Cornelius, P. L. and Anderson, R. L. Planned unbalanced designs for estimation of quantitative genetic parameters. I: Two-way matings. *Biometrics*, 1996, **52**, 56–70.
2. Kaart, T. Ülevaade geneetiliste parameetrite hindamisel kasutatavatest mudelitest. In *Eesti Põllumajandusülikooli Loomakasvatusinstituudi teadustöid*, **71** (Lokk, E., ed.). EPMÜ Loomakasvatusinstituut, Tartu, 2001, 52–67.
3. Visscher, P. M. On the sampling variance of intraclass correlations and genetic correlations. *Genetics*, 1998, **149**, 1605–1614.
4. Donner, A. and Koval, J. J. A note on the accuracy of Fisher's approximation to the large sample variance of an intraclass correlation. *Commun. Stat. Simul. Comput.*, 1983, **12**, 443–449.
5. Searle, S. R., Casella, G. and McCulloch, C. E. *Variance Components*. Wiley, New York, 1992.
6. McCulloch, C. E. and Searle, S. R. *Generalized, Linear and Mixed Linear Models*. Wiley, New York, 2001.
7. Khuri, A. I., Mathew, T. and Sinha, B. K. *Statistical Tests for Mixed Linear Models*. Wiley, New York, 1998.
8. Johnson, N. L. and Kotz, S. *Continuous Univariate Distributions – 2*. Wiley, New York, 1970.
9. Satterthwaite, F. E. Synthesis of variance. *Psychometrika*, 1941, **6**, 309–316.
10. Osborne, R. and Paterson, W. S. B. On the sampling variance of heritability estimates derived from variance analysis. *Proc. Roy. Soc. Edinburgh. Sec. B*, 1952, **64**, 456–461.
11. Swinger, L. A., Harvey, W. R., Everson, D. O. and Gregory, K. E. The variance of intraclass correlation involving groups with one observation. *Biometrics*, 1964, **20**, 818–826.
12. Ahrens, H. and Pincus, R. On two measures of unbalancedness in a one-way model and their relation to efficiency. *Biometrical J.*, 1981, **23**, 227–237.
13. SAS Institute Inc. *SAS OnlineDoc, Version 8*. Cary, NC, SAS Institute Inc., 1999.
14. Graybill, F. A. *Matrices with Applications in Statistics, Second Edition*. Wadsworth, Belmont, California, 1983.
15. Zerbe, G. O. and Goldgar, D. E. Comparison of intraclass correlation coefficients with the ratio of two independent F-statistics. *Commun. Stat. Theory Methods, Ser. A*, 1980, **9**, 1641–1655.
16. Hammarsley, J. M. The unbiased estimate and standard error of the interclass variance. *Metron*, 1949, **15**, 189–205.

Uus lähend dispersioonanalüüsiga hinnatud intraklass-korrelatsioonikordaja varieeruvusele

Tanel Kaart

On tuletatud valem populatsioonigeneetilistes uuringutes rakendatava peamise parameetri – intraklass-korrelatsioonikordaja – hinnangu dispersiooni arvutamiseks juhuslike mõjudega ühefaktorilisel dispersioonanalüüsil. Tuletatud valemi õigsust ja intraklass-korrelatsioonikordaja hinnangu täpsuse sõltuvust andmete struktuurist on uuritud modelleerimiseksperimentide abil. Lisaks on tõestatud, et intraklass-korrelatsioonikordaja hinnang on vähima varieeruvusega tasakaaluliste andmete korral.