

Kansas State University Libraries
New Prairie Press

Conference on Applied Statistics in Agriculture 2003 - 15th Annual Conference Proceedings

THE PROBABILITY OF PREPONDERANCY: AN ALTERNATIVE TO THE INTRACLASS CORRELATION

Ian R. Harris

Brent D. Burch

Follow this and additional works at: <https://newprairiepress.org/agstatconference>



Part of the [Agriculture Commons](#), and the [Applied Statistics Commons](#)



This work is licensed under a [Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License](#).

Recommended Citation

Harris, Ian R. and Burch, Brent D. (2003). "THE PROBABILITY OF PREPONDERANCY: AN ALTERNATIVE TO THE INTRACLASS CORRELATION," *Conference on Applied Statistics in Agriculture*. <https://doi.org/10.4148/2475-7772.1180>

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.

THE PROBABILITY OF PREPONDERANCY: AN ALTERNATIVE TO THE INTRACLASS CORRELATION

Ian R. Harris¹ and Brent D. Burch²

¹Department of Statistical Science
Southern Methodist University, Dallas, Texas 75275, U.S.A.

²Department of Mathematics and Statistics
Northern Arizona University, Flagstaff, Arizona 86011, U.S.A.

ABSTRACT

We propose a new parameter for measuring the influence of a random effect in a mixed linear model. This is the probability of preponderance of the random effect under study over the other random effects. In a one-way random effects model, this is simply the probability the group random effect is larger in absolute size than the individual random effect (or error). We discuss the meaning of the parameter and relate it to the more familiar intraclass correlation coefficient. The new parameter has the appealing property that it is applicable for any distribution, whereas the intraclass correlation has its origins in normally distributed random effects. Furthermore, the new parameter directly measures the random effect's impact on the observations whereas the intraclass correlation relies on the variances (second moments) of the random effects. We suggest parametric and nonparametric estimators of the parameter, and demonstrate the applicability of the results using real data. We also indicate how to extend the ideas to models with more than two sources of variation.

1 Introduction

Random and mixed effects models have a long tradition in statistical analysis, dating back to at least Airy (1861). Comprehensive reviews are given by Scheffé (1956) and Searle, Casella, and McCulloch (1992). Fisher (1918), defined variance as the square of the standard deviation, and proposed using proportions of total variance to describe the contribution of a particular effect. Later, (Fisher 1925) he pointed out that this ratio is the intraclass correlation, "...the correlation merely measures the relative importance of two groups of factors causing variation" (Fisher 1990, p. 223). Fisher (1918) specifically states that variance is to be used to measure variation as the data are usually to be taken to be normally distributed. This point is a primary motivation for the present paper. If the random effects are normally distributed then variance is clearly the correct measure of variation. However, if the random effects follow, for example, a Laplace (double exponential) distribution then one might argue that absolute deviations are more appropriate. In order to address this concern, we introduce a new parameter, which we term the probability of preponderancy, and denote it by θ .

The probability of preponderancy is the probability that the random effect under study

is larger in absolute size than the other random effects. Consider the one-way random effects model where the observation on the j^{th} individual in the i^{th} group, denoted by Y_{ij} , is

$$Y_{ij} = \mu + A_i + \epsilon_{ij}, \quad (1)$$

μ is the overall population mean, A_i is the group random effect, and ϵ_{ij} is the random error. The parameter under study is

$$\theta = P(|A_i| > |\epsilon_{ij}|). \quad (2)$$

The preponderance probability has a particularly easy interpretation as the proportion of individuals in which the group random effect is “more important”. For example, a familiar use of the model is to evaluate questions of genetic versus environmental effects on a phenotype. In such a model a θ of 0.3 could be interpreted as “in 30% of individuals the genetic effect is larger than the environmental effect”. As we will see in Section 3, for extreme values of the intraclass correlation ρ there are different qualitative impressions from ρ and θ . For example, a ρ of 0.01 would generally give the impression that genetic effects are almost negligible, but under normal distribution assumptions $\theta = 0.064$ and under Laplace distribution assumptions $\theta = 0.091$. Put another way, with $\rho = 0.01$ it is possible that in almost 10% of individuals, the group random effect is more important than the random error. A similar, but reversed pattern occurs for large ρ .

We suggest that the new parameter has an important role to play when measuring the influence of random effects. By definition, θ is the proportion of individuals for which the random effect is larger in absolute size than the error. This is in contrast to the intraclass correlation coefficient, where it is assumed that variance is the best measure of variation. The fact that variance may not be the best measure of variation is also a motivation behind the work of Cox and Hall (2002), who examine additive random effects and consider estimating cumulants of the distributions of the random effects.

The remainder of the paper is organized as follows. Section 2 briefly reviews the history of variance components and the intraclass correlation coefficient and also robustness of common estimators to nonnormality. Section 3 introduces the preponderance probability θ and considers its relationship to ρ . Section 4 discusses frequentist parametric inference about θ under normality assumptions. In Section 5 we introduce a nonparametric estimator based on a U -statistic. The estimator is consistent under certain conditions. We recommend using bootstrap methods to construct confidence intervals based on the nonparametric estimator, and discuss this in the section. An example is given in Section 6 which illustrates the usefulness of the parameter. There are some concluding comments in Section 7 and a brief discussion that extends the method to models with more than two sources of variation.

2 Background

According to Scheffé (1956) and Searle, Casella, and McCulloch (1992), the first explicit use of the one-way random effects model was made by Airy (1861). Fisher (1918, 1925)

used a two variance component model in quantitative genetics by considering additive genetic effect as one random effect and lumping together non-additive genetic effects with environmental effects to form a second random effect. See Bennett (1983) for details. Fisher specifically advocated using variances to describe variation as errors were to be taken to be normally distributed.

The intraclass correlation coefficient, first named by Fisher (1925) is generally defined in the terms of variance components and assumes normally distributed random effects. For a thorough overview of the intraclass correlation coefficient in the one-way random effects model, see Donner (1986). Despite the clear dominance in the literature of the assumptions of normality and using variance to measure variation, there are some papers where the intraclass correlation coefficient is not based on these assumptions. Rothery (1979) defines a nonparametric measure of intraclass correlation as a probability of certain types of concordances among the observations. The estimator is a function of the ranks of the observations. Shirahata (1982) compares the estimator derived by Rothery (1979) to two additional nonparametric estimators. Commenges and Jacqmin (1994) provide a definition of the intraclass correlation coefficient which depends on variances but not normal theory. The parameter is a function of the conditional expectation of the observations given the random effect under study. The statistic used to test the hypothesis of null intraclass correlation is related to the pairwise correlation coefficient, which predates the ANOVA estimator of ρ . See Karlin, Cameron, and Williams (1981) for a general class of weighted pairwise correlation coefficients. Bansal and Bhandary (1994) discuss the properties of M-estimators of the intraclass correlation coefficient. Müller and Büttner (1994) note that selecting the appropriate estimator of ρ depends on the underlying sampling theory. Vogler, Wette, McGue, and Rao (1995) compare estimators of ρ under a variety of sampling conditions. Cox and Hall (2002) propose a linear model, but do not assume the random effects are normally distributed. Instead of looking at variances, they estimate cumulants of the distributions of the random effects in order to compare their relative importance.

It is our intention to use an underlying linear model structure, that is, consider scenarios in which one can identify sources that produce variation in the observations, and where the sources combine additively. Furthermore, the parameter of interest should be a direct function of the sources, not the variances (or second moments) of the sources. The papers mentioned above do not address these issues.

3 The probability of preponderancy

Consider the one-way random effects model given by

$$Y_{ij} = \mu + A_i + \epsilon_{ij}, \quad (3)$$

where $i = 1, \dots, a$, $j = 1, \dots, b_i$, and $\sum_{i=1}^a b_i = n$. Y_{ij} is the j^{th} observation in the i^{th} class (or group) of factor A . The a groups of A in the model are assumed to be randomly selected from some large population of groups. Furthermore, a random sample of size b_i has been

obtained from the i^{th} group. ϵ_{ij} is referred to as random error. It is often assumed that the A_i are iid with mean 0 and variance σ_1^2 , the ϵ_{ij} are iid with mean 0 and variance σ_2^2 , and that A_i and ϵ_{ij} are mutually independent, with $\sigma_1^2 \geq 0$ and $\sigma_2^2 > 0$. In Section 4 of this paper we will assume the random effects are normally distributed, but in much of the paper this is not the case. The parameter μ is a fixed but unknown quantity that represents the overall mean of Y_{ij} .

Observations within the same group are correlated ($Cov(Y_{ij}, Y_{ij'}) = \sigma_1^2$) and observations from different groups are uncorrelated. In addition, $Var(Y_{ij}) = Var(A_i) + Var(\epsilon_{ij}) = \sigma_1^2 + \sigma_2^2$. The intraclass correlation coefficient is $\rho = \sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$ where $0 \leq \rho < 1$. ρ may be interpreted as the proportion of the variation in the Y_{ij} 's attributed to factor A . Note that ρ is a function of the variances of the random effects; and is not a direct comparison of random effects themselves.

The preponderance probability, θ , is defined as

$$\theta = P(|A_i| > |\epsilon_{ij}|). \quad (4)$$

This probability has a direct interpretation in terms of random effects. For example, in genetic applications it can be interpreted as the probability that additive genetic effects are more important than "other" effects for the phenotype under investigation. Note that θ directly contrasts the influence of A on Y with the influence of ϵ on Y .

If the random effects are normally distributed, it can be shown that

$$\begin{aligned} \theta &= P(A_i^2 > \epsilon_{ij}^2) \\ &= P\left(\frac{\epsilon_{ij}^2/\sigma_2^2}{A_i^2/\sigma_1^2} < \sigma_1^2/\sigma_2^2\right) \\ &= F_{1,1}(\sigma_1^2/\sigma_2^2) \end{aligned} \quad (5)$$

where $F_{1,1}(\cdot)$ is the cumulative distribution function of an F -distributed variate having numerator and denominator degrees of freedom equal to one. By definition, $0 \leq \theta < 1$. In this case the parameter can be written as

$$\begin{aligned} \theta &= F_{1,1}(\rho/(1-\rho)) \\ &= \frac{2}{\pi} \tan^{-1} \sqrt{\frac{\rho}{1-\rho}} \\ &= \frac{2}{\pi} \sin^{-1} \sqrt{\rho} \end{aligned} \quad (6)$$

and thus

$$\rho = \sin^2\left(\frac{\pi}{2}\theta\right). \quad (7)$$

Note that when $\rho = 0$, $\theta = 0$, when $\rho = 1/2$, $\theta = 1/2$, and as ρ approaches one, θ approaches one.

One alternative distribution for the random effects is Laplace (double exponential). Suppose that A_i has a Laplace distribution with mean 0 and variance σ_1^2 , ϵ_{ij} is Laplace with mean 0 and variance σ_2^2 , and A_i and ϵ_{ij} are mutually independent. In this scenario it is known that $|A_i|$ and $|\epsilon_{ij}|$ have exponential distributions with means equal to $\sigma_1/\sqrt{2}$ and $\sigma_2/\sqrt{2}$, respectively. It can be shown that

$$\begin{aligned} \theta &= P(|A_i| > |\epsilon_{ij}|) \\ &= P\left(\frac{|\epsilon_{ij}|/\sigma_2}{|A_i|/\sigma_1} < \sigma_1/\sigma_2\right) \\ &= F_{2,2}(\sigma_1/\sigma_2) \end{aligned} \tag{8}$$

where $F_{2,2}(\cdot)$ is the cumulative distribution function of an F -distributed random variable having numerator and denominator degrees of freedom equal to two. Note that as in the normal case, by definition $0 \leq \theta < 1$. After some algebra, the parameter can also be written as

$$\theta = \frac{\rho^{1/2}}{(1 - \rho)^{1/2} + \rho^{1/2}}. \tag{9}$$

Alternatively,

$$\rho = \frac{\theta^2}{(1 - \theta)^2 + \theta^2}. \tag{10}$$

As in the normal case, when $\rho = 0$, $\theta = 0$, when $\rho = 1/2$, $\theta = 1/2$, and as ρ approaches one, θ approaches one. For other values of ρ , the relationship between ρ and θ is somewhat different depending on which of these two distributional forms we adopt. Table 1 displays the values of θ as a function of $0 \leq \rho < 0.5$ for the normal and Laplace distributional assumptions.

One can make some general statements about the relationship between θ and ρ , at least in the case of continuous distributions with finite second moments. Note that $\rho = 0$ iff $\sigma_1^2 = 0$ and $\sigma_2^2 > 0$. Of course $\sigma_1^2 = 0$ iff $A = 0$ almost everywhere. Since ϵ is continuous with positive variance, we then have that these conditions are satisfied iff $P(|A| > |\epsilon|) = 0$, i.e., $\theta = 0$. Similar reasoning also establishes that ρ approaches 1 iff θ approaches 1. It seems reasonable, and we conjecture that $\rho > 1/2$ iff $\theta > 1/2$, but we as yet have no proof of this fact. If the variables in question are discrete then it is possible for other relationships to occur, for example, ρ can be very small, but θ be quite large.

It is important to note that θ and ρ measure different phenomena. θ relates to individuals, telling us for what proportion of individuals, the group random effect is “more important” than the individual random effect. ρ tells us what proportion of the variance of a trait in a population is due to the group effect. Using variance to measure variation is a concept tied to normal theory. It follows that using ρ to measure the relative contribution of the random effect to the variation in the observations may be misleading. For

example, one might think that $\rho = 0.01$ would imply that group effects are negligible, but under the assumption of Laplace distributions, in over 9% of individuals, the group effect outweighs the individual effect. Again, $\rho = 0.1$ would usually be taken to be a small group effect, but under Laplace distributions, $\theta = 0.25$, so that for 25% of individuals, the group effect dominates. We offer θ as an alternative to ρ and suggest its appropriateness is not dependent on a specific distribution.

In the next two sections we take up the problem of inference from data, first under parametric, and then under nonparametric assumptions.

4 Inference under normal distribution assumptions

Although the main intent of this paper is to introduce a parameter and associated estimator that do not assume normality, it is still useful to look at inference about the parameter when normality does hold. Assume that A_i is $N(0, \sigma_1^2)$ and ϵ_{ij} is $N(0, \sigma_2^2)$, with the usual independence assumptions. A commonly chosen estimator of ρ is the restricted maximum likelihood (REML) estimator, which we denote by $\hat{\rho}$. See Searle, Casella, and McCulloch (1992, pages 90ff, 159ff, 249ff) for a general description of REML estimators of variance components. Due to the invariance of maximum likelihood estimators, the REML estimator of θ is

$$\hat{\theta} = \frac{2}{\pi} \sin^{-1} \sqrt{\hat{\rho}}. \quad (11)$$

Maximum likelihood estimators, ANOVA estimators, and pairwise estimators of ρ are also discussed by Vogler et al. (1995).

The analysis of variance table for the balanced one-way random effects model is given in Table 2. For balanced data, where $b_i = b$ for $i = 1, \dots, a$, the REML estimator of ρ is the same as the ANOVA estimator of ρ bound to the parameter space. In these scenarios an exact confidence interval for θ is readily available. It is well known that Q_1 (the between group sum of squares) and Q_2 (the within group or error sum of squares) are independently distributed χ^2 variables and so

$$\left(1 + b \frac{\sigma_1^2}{\sigma_2^2} \right) \frac{(a-1)Q_1}{a(b-1)Q_2} \sim F(a(b-1), a-1). \quad (12)$$

Let $F_{\alpha/2}$ and $F_{1-\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ percentiles of the F distribution having numerator and denominator degrees of freedom equal to $a(b-1)$ and $a-1$, respectively. A $100(1-\alpha)\%$ equal-tailed confidence interval for θ is obtained by recognizing that

$$\begin{aligned} 1 - \alpha &= P \left(F_{\alpha/2} < \left(1 + b \frac{\sigma_1^2}{\sigma_2^2} \right) \frac{(a-1)Q_1}{a(b-1)Q_2} < F_{1-\alpha/2} \right) \\ &= P \left(\frac{1}{b} \left(F_{\alpha/2} \frac{a(b-1)Q_2}{(a-1)Q_1} - 1 \right) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{b} \left(F_{1-\alpha/2} \frac{a(b-1)Q_2}{(a-1)Q_1} - 1 \right) \right) \end{aligned}$$

$$= P \left(F_{1,1} \left[\frac{1}{b} \left(F_{\alpha/2} \frac{a(b-1)Q_2}{(a-1)Q_1} - 1 \right) \right] < \theta < F_{1,1} \left[\frac{1}{b} \left(F_{1-\alpha/2} \frac{a(b-1)Q_2}{(a-1)Q_1} - 1 \right) \right] \right) \quad (13)$$

since $F_{1,1}(\cdot)$ is a monotone increasing function. Simplifying, a $100(1-\alpha)\%$ equal-tailed confidence interval for θ is

$$\left(\frac{2}{\pi} \tan^{-1} \sqrt{\frac{1}{b} \left(F_{\alpha/2} \frac{a(b-1)Q_2}{(a-1)Q_1} - 1 \right)}, \frac{2}{\pi} \tan^{-1} \sqrt{\frac{1}{b} \left(F_{1-\alpha/2} \frac{a(b-1)Q_2}{(a-1)Q_1} - 1 \right)} \right). \quad (14)$$

The asymptotic properties of the REML estimator of θ follow from the asymptotic properties of $\hat{\rho}$. Burch and Harris (2001) show that

$$\hat{\rho} \stackrel{asympt}{\sim} N(\rho, V(\hat{\rho})) \quad (15)$$

where the asymptotic variance of $\hat{\rho}$ given by

$$V(\hat{\rho}) = \frac{2(n-1)(1-\rho)^2 (1+\rho(b-1))^2}{(n-a)(a-1)b^2}. \quad (16)$$

This formula was first derived by Fisher (1925). Using a Taylor expansion approach, it is well known that if a function $g(\cdot)$ is differentiable at ρ and $V(\hat{\rho})$ goes to zero as the sample size increases, then

$$\hat{\theta} \stackrel{asympt}{\sim} N(g(\rho), (g'(\rho))^2 V(\hat{\rho})). \quad (17)$$

In our case, $g(\rho) = 2/\pi \sin^{-1} \sqrt{\rho}$ and the asymptotic distribution $\hat{\theta}$ can be expressed of terms of ρ as

$$\hat{\theta} \stackrel{asympt}{\sim} N \left(\frac{2}{\pi} \sin^{-1} \sqrt{\rho}, \frac{2(n-1)}{\pi^2(n-a)(a-1)} \frac{(1-\rho)(1+\rho(b-1))^2}{b^2 \rho} \right) \quad (18)$$

and in terms of θ as

$$\hat{\theta} \stackrel{asympt}{\sim} N \left(\theta, \frac{2(n-1)}{\pi^2(n-a)(a-1)} \frac{(1 + \sin^2(\frac{\pi}{2}\theta)(b-1))^2}{b^2 \tan^2(\frac{\pi}{2}\theta)} \right). \quad (19)$$

5 Nonparametric inference

For maximum utility, inference about θ should be robust to choice of the parametric family. One way to achieve this is to use nonparametric methods. In this section it is assumed that the A_i are iid with mean 0 and variance σ_1^2 , the ϵ_{ij} are iid with mean 0 and variance σ_2^2 , and that A_i and ϵ_{ij} are mutually independent with unspecified distributions. For simplicity we will assume a balanced design. Along with a nonparametric point estimator, we

construct a nonparametric confidence interval for θ using a bootstrap sampling procedure for hierarchical data.

We begin by recognizing that each of the a levels of A may be compared with each of the ab values of ϵ to form a U -statistic. Specifically,

$$\tilde{\theta} = \frac{1}{a^2b} \sum_{k=1}^a \sum_{i=1}^a \sum_{j=1}^b I(|A_k| > |\epsilon_{ij}|) \quad (20)$$

where $I(\cdot)$ is an indicator function that takes the value 1 if the condition is true and zero if not. Clearly, $\tilde{\theta}$ can never be calculated as it depends on the actual values of the random effects. Nevertheless, it is instructive to explore the properties of $\tilde{\theta}$.

Theorem: The estimator $\tilde{\theta}$ is consistent for θ and has an asymptotic normal distribution.

Proof: See technical report by Authors at <http://faculty.smu.edu/iharris/pubs.htm>.

Of course, we cannot observe $\tilde{\theta}$, and instead must work with estimated values of A and ϵ . This introduces considerable complication to the theory. Simple, or naive estimators of A and ϵ are

$$\hat{A}_i^{naive} = \bar{Y}_i - \bar{Y}_{..} \quad (21)$$

$$\hat{\epsilon}_{ij}^{naive} = Y_{ij} - \bar{Y}_{i.} \quad (22)$$

Using these, we have that

$$\hat{A}_i^{naive} = A_i - \bar{A}_{..} + \bar{\epsilon}_i - \bar{\epsilon}_{..} \quad (23)$$

$$\hat{\epsilon}_{ij}^{naive} = \epsilon_{ij} - \bar{\epsilon}_i \quad (24)$$

and so one can propose using

$$\hat{\theta}^{naive} = \frac{1}{a^2b} \sum_{k=1}^a \sum_{i=1}^a \sum_{j=1}^b I(|\hat{A}_k^{naive}| > |\hat{\epsilon}_{ij}^{naive}|). \quad (25)$$

The presence of the extra terms induces dependence among the \hat{A} 's, and the distribution of \hat{A}_i depends on the distribution of ϵ . These dependences only disappear if we have $b \rightarrow \infty$ as well as $a \rightarrow \infty$ as outlined in the previously mentioned technical report. From a practical viewpoint this might seem problematic, as b is usually small. However, simulation results outlined later are encouraging for $b \geq 5$. While it is true that \hat{A}_i^{naive} and $\hat{\epsilon}_{ij}^{naive}$ play the role of A_i and ϵ_{ij} in that $E(\hat{A}_i^{naive}) = E(A_i)$ and $E(\hat{\epsilon}_{ij}^{naive}) = E(\epsilon_{ij})$, the variances of \hat{A}_i^{naive} and $\hat{\epsilon}_{ij}^{naive}$ do not equal the variances of A_i and ϵ_{ij} .

We also consider an estimator of A based on jackknife versions of Q_1 and Q_2 . That is, determine $Q_{1(-i)}$ and $Q_{2(-i)}$, where $(-i)$ denotes that Q_1 and Q_2 are computed by excluding the observations in the i^{th} group. It can be shown that

$$Var \left(\sqrt{\frac{a}{a-1} \left(1 - \frac{(a-4)Q_{1(-i)}}{(a-1)(b-1)Q_{2(-i)}} \right)} (\bar{Y}_{i.} - \bar{Y}_{..}) \right) = \sigma_1^2. \quad (26)$$

For details of the derivation see the technical report. Note that we require $a > 4$. An alternative estimator of A to consider is

$$\hat{A}_i^{jackknife} = \sqrt{\max \left\{ 0, \frac{a}{a-1} \left(1 - \frac{(a-4)Q_{1(-i)}}{(a-1)(b-1)Q_{2(-i)}} \right) \right\}} (\bar{Y}_i - \bar{Y}_{..}) \quad (27)$$

Since the argument of the square root and $\bar{Y}_i - \bar{Y}_{..}$ are uncorrelated, $E(\hat{A}_i^{jackknife}) = E(A_i)$ and $Var(\hat{A}_i^{jackknife})$ should be close to $Var(A_i)$. We use $\hat{\theta}_{jackknife}$ to denote the estimator of θ based on $\hat{A}_i^{jackknife}$ and $\hat{\epsilon}_{ij}^{scalar} = \sqrt{b/(b-1)}(Y_{ij} - \bar{Y}_i)$ since

$$Var(\hat{\epsilon}_{ij}^{scalar}) = \sigma_2^2 \quad (28)$$

We now evaluate the practical use of the estimators of θ . Using normal or Laplace distribution assumptions, simulation results suggest no estimator of the form $\hat{\theta}$ uniformly outperforms another estimator in terms of MSE for the values of a and b under consideration. For instance, $\hat{\theta}_{naive}$ exhibits poor performance when θ is small due to large biases. However, $MSE(\hat{\theta}_{naive})$ is comparable to the MSE of the other estimator when the parameter is large. To select the appropriate estimator, we consider the simulated coverage probabilities of the two competing nonparametric estimators ($\hat{\theta}_{naive}$ and $\hat{\theta}_{jackknife}$) using bootstrap confidence intervals.

Efron and Tibshirani (1998) discuss nonparametric bootstrap estimation techniques when resampling is based on nonhierarchical data. The natural layering or nested feature of the data in the problems we consider present a complication when compared to the usual bootstrap resampling methods. Davison and Hinkley (1997, p.100-102) provide an outline of the resampling procedure for hierarchical data having two stages of sampling. The strategy recommended by Davison and Hinkley (1997) attempts to match the resampling variations of the statistics to the variational properties of the data. See the technical report for further details.

A simulation study was conducted to evaluate the performance of the bias-corrected (BC) bootstrap confidence intervals using the naive and jackknife forms of the estimators. Performance was judged by the simulated coverage probability. For $a = 10$, $b = 5$, 90% BC confidence intervals were built from normal, Laplace, and uniform distributed data for $\theta = 0.1, 0.5$, and 0.9 . Simulation results are displayed in Table 3. In general, the coverage probabilities associated with the jackknife estimator are more consistent. This suggests that $\hat{\theta}_{jackknife}$ is preferred over the naive estimator. A more comprehensive simulation study evaluating coverage probabilities for a different values of a , b , ρ , and a variety of distributions may yield additional information.

6 Example

Gibbons and Bhaumik (2001) compared the results of copper concentrations as determined by seven laboratories. Water samples, prepared by an independent source, were analyzed

by the laboratories in a blind interlaboratory study. For brevity reasons, we consider that part of the dataset consisting of five replications per laboratory based on a copper concentration of zero $\mu\text{g/L}$. The copper concentrations as measured by the laboratories are given in Table 4. Negative values are possible since the copper concentrations are based on a linear calibration function.

Using a one-way random effects model, one can determine how the variability of the laboratories contributes to the overall variability of the measurements. If σ_1^2 represents the variance of the laboratories and σ_2^2 represents the variance of the measurements within laboratories (or error), the intraclass correlation coefficient is $\rho = \sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$. For this data, $Q_1 = 230.32$, $Q_2 = 60.08$, and $\hat{\rho} = 0.04$. That is, the percentage of variance in copper concentrations associated with the variance in laboratories is 4%.

Recall that the probability of preponderancy (θ) and ρ measure different things. θ relates to individual measurements, indicating the proportion of individuals for which the laboratory effect is “more important” than the error effect. Under normal distribution assumptions, $\hat{\theta} = 0.13$ and a 95% confidence interval for θ is (0.00, 0.51). The value of 0.13 should be interpreted as follows; “in 13% of determinations, the laboratory effect is greater than individual measurement error”. In this application we find that the intraclass correlation coefficient does not clearly indicate the impact laboratories have on the copper concentration measurements. The nonparametric estimator using the naive approach is 0.43 and using the jackknife approach is 0.24.

7 Some extensions

It is possible to extend the ideas to models with more than one component. For example suppose $Y = \mu + A + B + \epsilon$. Without worrying too much about indices, or whether we have crossing or nesting, there are at least two possible extensions to θ . The first is

$$\theta_A = P(|A| > \max(|B|, |\epsilon|)) \tag{29}$$

the second is

$$\lambda_A = P(|A| > |B + \epsilon|). \tag{30}$$

The first attempts to find for what proportion of individuals is A the greatest of the three effects. The second finds for what proportion of individuals is A greater than all other effects combined. Similar expressions could be formed for B . Under normal theory, with A having variance σ_1^2 , B having variance σ_2^2 , and ϵ with variance σ_3^2 , then one can show that

$$\lambda_A = F_{1,1} \left(\frac{\sigma_1^2}{\sigma_2^2 + \sigma_3^2} \right). \tag{31}$$

8 Summary and conclusions

We have presented a new parameter, which we call the probability of preponderancy. In the one-way random effects model, this probability is the proportion of individuals for whom the group random effect is larger than the individual random effect. In genetic applications this could be interpreted as the proportion of individuals for whom the genetic effect dominates environmental effects. This parameter is an appealing complement to the more familiar intraclass correlation coefficient since its relevancy does not hinge on a particular distribution. We have presented both parametric and nonparametric estimators. Bias-corrected bootstrap confidence intervals associated with the nonparametric estimators may be employed when $a \geq 10, b \geq 5$. The estimator incorporating a jackknife approximation to a scalar exhibits the most consistent coverage probabilities. Actual confidence levels depend on the underlying distributions.

Ideas for the extension of probability of preponderancy to models with more than one component of variation were briefly discussed.

REFERENCES

- AIRY, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combinations of Observations*. London: MacMillan.
- BANSAL, N. K. & BHANDARY, M. (1994). Robust M-estimation of the intraclass correlation coefficient. *Australian Journal of Statistics* **36**, 287–301.
- BURCH, B. D. & HARRIS, I. R. (2001). Closed-form approximations to the REML estimator of a variance ratio (or heritability) in a mixed linear model. *Biometrics* **57**, 1148–1156.
- BENNETT, J. H. (1983). *Natural selection, heredity, and eugenics: including selected correspondence of R.A. Fisher with Leonard Darwin and others*. Oxford: Clarendon Press.
- COMMENGES, D. & JACQMIN, H. (1994). The intraclass correlation coefficient: distribution-free definition and test. *Biometrics* **50**, 517–526.
- COX, D. R. & HALL, P. (2002). Estimation in a simple random effects model with non-normal distributions. *Biometrika* **89**, 831–840.
- DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.
- DONNER, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review* **54**, 67–82.

- EFRON, B. & TIBSHIRANI, R. J. (1998). *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press LLC.
- FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52**, 399–433.
- FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.
- FISHER, R. A. (1990). *Statistical Methods, Experimental Design and Scientific Inference*. Oxford Press.
- GIBBONS, R. D. & BHAUMIK, D. K. (2001). Weighted random-effects regression models with application to interlaboratory calibration. *Technometrics* **43**, 192–198.
- KARLIN, S., CAMERON, E. C. & WILLIAMS, P. T. (1981). Sibling and parent-offspring correlation estimation with variable family size. *Proceedings of the National Academy of Sciences, U.S.A.* **78**, 2664–2668.
- MÜLLER, R. & BÜTTNER, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine* **13**, 2465–2476.
- ROTHERY, P. (1979). A nonparametric measure of intraclass correlation. *Biometrika* **66**, 629–639.
- SEARLE, S. R., CASELLA, G. & MCCULLOCH, C. E. (1992). *Variance Components*. New York: John Wiley.
- SCHEFFÉ, H. (1956). Alternative models for the analysis of variance. *The Annals of Mathematical Statistics* **27**, 251–271.
- SHIRAHATA, S. (1982). Nonparametric measures of intraclass correlation. *Communications in Statistics - Theory and Methods* **11**, 1707–1721.
- VOGLER, G. P., WETTE, R., MCGUE, M. K. & RAO, D. C. (1995). Properties of alternative estimators of familial correlations under variable sibship size. *Biometrics* **51**, 276–283.

Table 1: Relationship between ρ and θ

ρ	θ (Normal)	θ (Laplace)
.01	.064	.091
.05	.144	.187
.10	.205	.250
.15	.253	.296
.20	.295	.333
.25	.333	.366
.30	.369	.396
.35	.403	.423
.40	.436	.449
.45	.468	.475
.50	.500	.500

Table 2: ANOVA Table

Source	df	Sum of Squares
Between Groups	$a - 1$	$Q_2 = \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{i.} - \bar{Y}_{..})^2$
Within Groups	$a(b - 1)$	$Q_1 = \sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{i.})^2$
Total	$ab - 1$	$\sum_{i=1}^a \sum_{j=1}^b (Y_{ij} - \bar{Y}_{..})^2$

Table 3: Comparison of Simulated Coverage Probabilities when $1-\alpha = 0.90$

θ	Distribution	Naive	Jackknife
0.1	Normal	0.42	0.80
	Laplace	0.12	0.82
	Uniform	0.76	0.81
0.5	Normal	0.94	0.93
	Laplace	0.94	0.94
	Uniform	0.93	0.90
0.9	Normal	0.89	0.88
	Laplace	0.87	0.86
	Uniform	0.89	0.89

Table 4: Copper Concentrations ($\mu\text{g/L}$)

Lab 1	Lab 2	Lab 3	Lab 4	Lab 5	Lab 6	Lab 7
3.000	2.100	0.800	1.661	0.090	7.226	0.018
2.000	0.300	-0.185	1.996	-2.510	-1.000	-3.000
-1.000	2.000	0.990	0.000	7.270	0.000	0.000
1.000	1.300	0.905	2.993	7.140	10.244	-2.000
-1.000	2.000	0.365	2.042	0.280	-2.177	-2.000