Submitted to the Brazilian Journal of Probability and Statistics

Improved U-tests for variance components in one-way random effects models

Juvêncio Santos Nobre[a,1], **Julio M. Singer**[b] **and Maria J. Batista**[a]

[a] *Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Brazil*

[b] *Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil.*

E-mail: juvencio@ufc.br ; jmsinger@ime.usp.br ; jacqueline@ufc.br

Abstract. Based on a decomposition of a U-statistic, Nobre et al. (2008, Festschrift to P.K. Sen, *IMS Lectures Notes Monograph Series*) proposed a test for the hypothesis that the within-treatment variance component in a one-way random effects model is null, specially useful when very mild assumptions are imposed on the underlying distributions. We consider a bootstrap version of that U-test and evaluate its performance via simulation studies in different scenarios. The bootstrap U-test has better statistical properties than the original test even in small samples. Furthermore, it is easy to implement and has a low computational cost. We consider two examples with unbalanced small sample datasets, for illustrative purposes.

1 Introduction

Consider the one-way random effects model

$$y_{ij} = \mu + b_i + e_{ij},\ i = 1, \dots k,\ j = 1, \dots, n_i\ (\geq 2) \tag{1}$$

with b_i and e_{ij} denoting independent random variables with null means and variances σ_b^2 and σ_e^2, respectively. The parameter μ is the mean response, b_i represents the random effect associated to the i-th treatment and e_{ij} represents a random measurement error associated with the j-th observation obtained under the i-th treatment. Here, σ_b^2 and σ_e^2 are the between- and within-treatment variance components, respectively.

In general, data analysis based on such a model focuses on the estimation of μ and on testing

[1] Corresponding author. Universidade Federal do Ceará, Campus do Pici, Bloco 910, Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Brazil, CEP: 60440-900.

MSC 2010 subject classifications: 62G86, 62F03, 62F30, 62F40

Keywords and phrases. Bootstrap, martingales, nonstandard hypothesis, one-way random effects model, U-statistics

the hypothesis of no treatment effects, namely

$$\mathcal{H}_0 : \sigma_b^2 = 0 \quad \text{vs} \quad \mathcal{H}_1 : \sigma_b^2 > 0. \tag{2}$$

Inference about variance components in random effects models and more specifically in linear mixed models, has a long history in the statistical literature. In this context, McCulloch et al. (2008) provide an excellent overview of estimation and prediction while Khuri et al. (1998) and Demidenko (2013) present extensive reviews of this topic. With exception of some special situations (under the assumption of normality) there are no exact tests for the hypothesis of null variance components, as discussed in Khuri et al. (1998), Lencina et al. (2005) and Demidenko (2013), for example. Asymptotic tests are needed in more general situations as we will discuss.

Under the additional assumption that b_i and e_{ij} follow normal distributions, the usual F-statistic for testing (2) is

$$F = \frac{\text{SQ}(b)/(k-1)}{\text{SQ}(e)/(n-k)}, \tag{3}$$

where $\text{SQ}(b) = \sum_{i=1}^{k} n_i(\overline{y}_{i.} - \overline{y}_{..})^2$ and $\text{SQ}(e) = \sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} n_i \overline{y}_{i.}^2$ are, respectively, the between- and within-treatment sums of squares with dots indicating over which indices the averages are computed. The F-statistic (3) follows a central F distribution with $k-1$ and $n-k$ degrees of freedom when $\mathcal{H}_0 : \sigma_b^2 = 0$ is true. In the balanced case (*i.e.*, when $n_1 = \ldots = n_k$) the test is uniformly most powerful invariant (UMPI). This optimality property does not hold in the unbalanced case. Details may be found in Khuri et al. (1998), for example.

Nobre (2007) and Nobre et al. (2008) provide an alternative test based on the decomposition of U-statistics in a nonparametric setup. Although it is not an exact test, it has good properties for moderate sample sizes and does not require the normality assumption. The test is derived under the assumption that $\mathbb{E}[e_{ij}^4] < \infty$ and thus accommodates a large class of distributions (not necessarily absolutely continuous) underlying the source of variation in model (1). The proposed test is also valid for other situations, like for tests of null variance components in heteroskedastic random effects as discussed in Nobre et al. (2008).

The class of U-statistics has its genesis in the papers of Halmos (1946) and Hoeffding (1948) and is well known for its simple structure and for the weak assumptions required for its use in statistical inference. It also provides a unified paradigm in the field of nonparametric Statistics and has been used in many applications, as illustrated in Lee (1990), Kowalski et al. (2002), Schaid et al. (2005), Sen (2006), Kowalski and Tu (2007), Nobre et al. (2008), Pinheiro et al. (2009) and

Nobre et al. (2013), among others. The related theory is available in many sources, among which we mention Serfling (1980), Sen (1981), Lee (1990) or Sen et al. (2010), for example.

In the context under investigation, the derivation of tests for (2) may not follow the standard procedures since the null hypothesis defines a point (or region) on the boundary of the parameter space and this brings in some technical difficulties. Asymptotic tests for (2) or, more generally, for testing the significance of variance components under linear mixed models are available in the literature. Based on the ideas of Silvapulle and Silvapulle (1995), Verbeke and Molenberghs (2003) obtained score-type tests under the assumption that the underlying probability distributions are normal. Along the same lines, Savalli et al. (2006) extended the results to accommodate elliptical underlying distributions. In particular, for the one-way random effects models, the corresponding test statistic follows an asymptotic distribution given by a 50:50 mixture of χ^2_0 and χ^2_1 distributions. Tests based on generalized likelihood methods (that are asymptotically equivalent to score-type tests) are considered in Self and Liang (1987), Stram and Lee (1994) and Silvapulle and Sen (2005), for example. The main disadvantage of such tests is the difficulty in verifying the required regularity conditions as shown in Giampaoli and Singer (2009). Other alternatives have been suggested in the literature as in Lin (1997), Hall and Praestgaard (2001), Zhu and Fung (2004), Zhang and Lin (2008), Crainiceanu and Ruppert (2004), Crainiceanu (2008), Greven et al. (2008) and Sinha (2009). In practice, all these results are difficult to apply, specially when the dimension of the vector of random effects is large; furthermore, they are only valid for some classes of distributions. The derivation of the proposed U-test is not affected by such difficulties and we envisage that it may serve as a building block for more general setups, as indicated in Nobre (2007) and Nobre et al. (2008, 2013).

Although there exists an exact F-test with optimal properties for testing wheteher the between-treatments variance component is null in a one-way random effects model with balanced data under normality, we must rely on sub-optimal or approximate tests in unbalanced or nonnormal settings. The asymptotic U-test that may be employed with unbalanced data and does not require a specified form for the underlying distributions. Nobre et al. (2008) advocate that to test $\mathcal{H}_0$ in situations where the distribution of the random effects and within-treatment errors are nonnormal, the U-test is preferable even when the number of treatments is small. Simulation studies indicated that the U-test is more powerful than the F-test, mainly for small and moderate samples. However, for small samples, the U-test is very liberal, that is, the size of the test (the true probability of falsely rejecting the null hypothesis) is greater than the nominal significance level. To bypass this problem, we obtain the empirical distribution of the test statistic for $\mathcal{H}_0$ via bootstrap methods.

This generates an exact test for (2) that does not depend on the normality of the b_i and e_{ij}. The key idea is to resample via the fitted model to create replicate datasets with the objective of obtaining the exact (empirical) distribution of the U-test proposed by Nobre et al. (2008).

In Section 2, we summarize the decomposition of the U-statistic that underlies the test as in Nobre et al. (2008). In Sections 3 and 4 we present the parametric and nonparametric versions of the bootstrap U-test and present a simulation study to evaluate its properties. In Section 5 we apply the proposed bootstrap test to small sample unbalanced nonnormal data. We conclude in Section 6 with a brief discussion and future research proposals.

2 Outline of the U-test

Consider the one-way random effects model (1) and suppose that the focus is on testing the hypothesis in (2). Let $g(x,y) = (x-y)^2/2$ and note that, under model (1), $\mathbb{E}[g(y_{ij}, y_{ij'})] = \mathbb{E}[(e_{ij} - e_{ij'})^2]/2 = \sigma_e^2$. An unbiased estimator of σ_e^2, based only on the n_i observations obtained under the i-th treatment ($i = 1, \ldots, k$) is given by the following U-statistic

$$U_i = \binom{n_i}{2}^{-1} \sum_{1 \le j < j' \le n_i} g(y_{ij}, y_{ij'}) = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i.})^2 = S_i^2. \tag{4}$$

Since $\mathbb{E}[(b_i - b_{i'})(e_{ij} - e_{i'j'})] = 0$, it follows that $\mathbb{E}[g(y_{ij}, y_{i'j'})] = \{2\sigma_b^2 + 2\sigma_e^2\}/2 = \sigma_b^2 + \sigma_e^2$. Therefore, an unbiased estimator of $\sigma_b^2 + \sigma_e^2$, based only on the observations obtained under treatments i and i' is given by the following generalized U-statistic of order (1,1)

$$U_{ii'} = (n_i n_{i'})^{-1} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} \frac{(y_{ij} - y_{i'j'})^2}{2}, \;\; 1 \le i < i' \le k. \tag{5}$$

Letting $n = \sum_{i=1}^{k} n_i$, the lexicographically ordered observations,

$$y_{11}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{2n_2}, \ldots, y_{k1}, \ldots, y_{kn_k},$$

may be re-expressed as

$$y_1, \ldots, y_{n_1}, y_{n_1+1}, \ldots, y_{n_1+n_2}, \ldots, y_{n_1+\ldots+n_{k-1}+1}, \ldots, y_n,$$

where the first n_1 observations relate to treatment 1, the next n_2, to treatment 2 and so on. The uniformly minimum variance unbiased estimator (UMVUE) of the variance of the observations is

given by the U-statistic

$$\begin{aligned} U_n^0 &= \binom{n}{2}^{-1} \sum_{1\le r<s\le n} \frac{1}{2}(y_r - y_s)^2 = \binom{n}{2}^{-1} \left\{ \sum_{i=1}^{k} \binom{n_i}{2} U_i + \sum_{1\le i<i'\le k} n_i n_{i'} U_{ii'} \right\} \\ &= \sum_{i=1}^{k} \frac{n_i(n_i-1)}{n(n-1)} U_i + 2 \sum_{1\le i<i'\le k} \frac{n_i n_{i'}}{n(n-1)} U_{ii'}, \end{aligned} \tag{6}$$

that is a linear combination of generalized U-statistics. The first and second terms in (6) correspond, respectively, to the within and between-treatment components. The U-statistic in (6) may be re-expressed as

$$U_n^0 = \sum_{i=1}^{k} \frac{n_i}{n} U_i + \sum_{1\le i<i'\le k} \frac{n_i n_{i'}}{n(n-1)} \{2U_{ii'} - U_i - U_{i'}\} = W_n + B_n, \tag{7}$$

where

$$W_n = \sum_{i=1}^{k} \frac{n_i}{n} U_i \text{ and } B_n = \sum_{1\le i<i'\le k} \frac{n_i n_{i'}}{n(n-1)} \{2U_{ii'} - U_i - U_{i'}\}.$$

Note that $\mathbb{E}[B_n] \ge 0$, so $\mathbb{E}[B_n] = 0$ if and only if $\sigma_b^2 = 0$. This fact motivated Nobre et al. (2008) to construct a test for (2) based on

$$B_n = \binom{n}{2}^{-1} \sum_{1\le r<s\le n} \eta_{nrs} \psi(y_r, y_s), \tag{8}$$

where

$$\eta_{nrs} = \begin{cases} \frac{n-n_i}{n_i-1} & \text{, if } y_r \text{ and } y_s \text{ are both observed under the } i\text{-th treatment} \\ -1 & \text{, otherwise.} \end{cases} \tag{9}$$

and $\psi(x_1, x_2) = (x_1 - \mu)(x_2 - \mu)$. Defining $M_n = \sum_{1\le r<s\le n} \eta_{nrs}^2$, Nobre et al. (2008), using the martingale property exhibited by B_n as demonstrated in Pinheiro et al. (2009) in a different setup, show that under $\mathcal{H}_0 : \sigma_b^2 = 0$,

$$J_n = \frac{\binom{n}{2} B_n}{W_n \sqrt{M_n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } k \to \infty. \tag{10}$$

Additionally, letting $\lim_{n\to\infty} M_n/n^3 = \lambda$, and assuming that the fourth moment of the distribution of the random effects is finite, they also showed that under the sequence of local hypotheses

$$\mathcal{H}_{1n} : \sigma_b^2 = \delta^2/\sqrt{n}, \tag{11}$$

it follows that

$$J_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(\frac{\delta^2}{2\sigma_e^2 \sqrt{\lambda}}, 1 \right) \text{ as } k \to \infty, \tag{12}$$

with δ representing a constant. Under the sequence $\mathcal{H}_{1n}$, the limiting normal distribution is shifted to the right by $\delta^2/(2\sigma_e^2\sqrt{\lambda})$. We may use J_n as a test statistic for (2), rejecting the null hypothesis $\mathcal{H}_0$ with significance level α when $J_n \geq z_\alpha$, where z_α represents the $(1-\alpha)100\%$ percentile of the standard normal distribution. By (12), the power of the test is directly related to the magnitude of the intraclass correlation coefficient $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$; more specifically, the power is a monotone increasing function of ρ, as expected.

These results are all asymptotic and may not necessarily be appropriate for samples of small/moderate sizes. In order to obtain a test with good properties even with small samples, we advocate using a bootstrap U-test, where the idea is to obtain the empirical distribution of the statistic J_n under $\mathcal{H}_0$ and use the fact that it suggests that the null hypothesis should be rejected for high values of J_n. In the next sections we discuss parametric and nonparametric bootstrap procedures. For the parametric bootstrap we will also study the effect of the misspecification of the conditional error distribution in order to evaluate the robustness of the method. To evaluate the possible effect of different distributions, we carry out simulations with asymmetric and heavy tailed distributions standardized (*i.e.*, with zero mean and variance 1) to make the results comparable.

3 A parametric bootstrap U-test

For the parametric bootstrap, we generate B pseudosamples under the null hypothesis as follows. Let

$$y^{(b_{ij})} = \widehat{\mu} + \widehat{\sigma}_e e_{ij},\ i = 1, \ldots k,\ j = 1, \ldots, n_i,\ b = 1, \ldots, B,$$

where e_{ij} represents a sequence of independent and identically distributed (*iid*) random variables with a given distribution. For each of $B = 999$ bootstrap pseudosamples, we obtain the statistics $J_1^*, J_2^*, \ldots, J_B^*$. Given the value of the statistic obtained from the original sample, J_n, the adjusted p-value for the bootstrap test (Davidson and Hinkley, 1997, p. 175) is

$$\widehat{p} = \frac{1 + \sum_{i=1}^{B} \mathbf{1}(J_i^* > J_n)}{B+1} \tag{13}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. To evaluate the behaviour of the proposed test for small and moderate samples we considered 10,000 Monte-Carlo samples obtained under model (1) with $\mu = 2$, $\sigma_e^2 = 1$, $b_i \sim \mathcal{N}(0, \sigma_b^2)$, for different distributions of e_{ij} and numbers of treatments (k = 5, 10, 30 and 100) in balanced studies. The within-treatment variance, σ_b^2, was set to 0 (to estimate the size of the test), 0.2, 0.5 or 1. The empirical power of the test under each

setting was evaluated for significance levels equal to 0.01, 0.05 and 0.10. Initially, we simulated data sets with different levels of imbalance, as in Nobre et al. (2008). Because the results did not differ too much from the balanced case, we only show those for the latter ($n_i = m, \forall i$). To evaluate the effect of misspecification of the generating distribution on the parametric procedure, the bootstrap pseudosamples were generated under normality assumptions even though the true underlying distribution was not normal.

We used both the REML estimator of σ_e obtained under of normality as well as the consistent estimator W_n. Here, also, the results were practically identical and for that reason, we only show the results based on W_n.

We repeated a similar simulation process considering $b_i \sim \{Y_i - \mathbb{E}[Y_i]\}/\sqrt{\text{Var}[Y_i]} \times \sigma_b$, where the *iid* random variables $Y_1, \ldots, Y_k$ follow a skew t distribution with 4.1 degrees of freedom, location parameter 0, dispersion parameter 1 and asymmetry parameter $\lambda = 1$ (St(0,1,1,4.1)) with index of skewness equal to 1.77. For details on the skew t distribution, see Azzalini and Capitanio (2003). The results were very similar and for that reason they were omitted. In Table 1 we show the results regarding the size and the empirical power of the test for different data generating distributions.

INSERT TABLE 1 HERE

The figures in Table 1 suggest that when the e_{ij} are normally distributed, the size of the bootstrap parametric U-test is very close to the nominal level, with a 10% maximum relative difference for the 1% significance level, less than 10% maximum relative difference for the 5% significance level and less than 4% maximum relative difference for the for 10% significance level, even with few treatments and few observations per treatment. The same conclusion holds when e_{ij} has an asymmetric distribution, with a small increase in the relative difference, principally for the 1% significance level. On the other hand, for heavy tailed distributions, the size of the bootstrap parametric U-test obtained under normality is not close to the nominal when there are few treatments and few repetitions per treatment ($k = 5$ and $m \leq 4$), mainly for the 1% significance level; in this case, the maximum relative difference increases to 20% in some settings. Otherwise, the results are quite satisfactory.

4 A nonparametric bootstrap U-test

In this context, we generated B pseudosamples under the null hypothesis, so that the observations $y_{ij}^{(b)}$, $b = 1, \ldots, B$ are randomly sampled from the set of the original observations $\{y_{ij}, i = 1, \ldots k,\ j = 1, \ldots, n_i\}$. For each of $B = 999$ bootstrap pseudosamples, we obtained the statistics

$J_1^*, J_2^*, \ldots, J_B^*$. Given the value of the statistic obtained from the original sample, J_n, the adjusted p-value is given as in (13).

Chernick and LaBudde (2011, Chapter 8) comment that bootstrap statistics may be inconsistent or unreliable; this occurs with the nonparametric bootstrap U-test mainly for the 1% significance level with $k = 5$ and $m = 2$, where the rejection rates are less than 50% of the nominal value. In a way, this is expected, since we are interested in an extreme case ($\alpha = 0.01$) and the sample sizes are small (only 10 observations). The same occurs when we consider few treatments. If we have at least 4 observations per treatment, the results displayed in Table 2 suggest that the size of the bootstrap nonparametric U-test is close to the nominal value, with maximum relative difference less than 10% for the 5% and 10% significance levels; a maximum relative difference of 20% (few cases) is observed for the 1% significance level, independently of the underlying distribution. As in the previous section, we repeated the simulation considering a standardized skew t distribution for the random effects obtaining very similar results (omitted here).

INSERT TABLE 2 HERE

5 Data examples

We consider an example originally presented in Snedecor and Cochran (1980) involving an experiment on artificial insemination of cows; several semen samples from a bull were tested for their ability to produce conceptions. The percentages of conceptions to services for successive samples from six randomly sampled bulls are displayed in Table 3.

INSERT TABLE 3 HERE

A model suggested by Alkhamisi (2000) is

$$y_{ij} = \mu + b_i + e_{ij},\ i = 1, \ldots 5,\ j = 1, \ldots, n_i$$

where y_{ij} represents the percentages of conceptions obtained from the jth sample taken from the ith bull, μ designates the overall mean, b_i designates the random effect due to the ith bull and e_{ij} denotes a random error. Given the bounded nature of the response variable, normality does not seem valid for either sources of variation. The objective of this example is to test whether the within-bull variance may be dropped from the model, that is, to test (2).

Crainiceanu and Ruppert (2004) show that for testing that a variance component in a one-way random effects model is null, the reference chi-bar-squared distributed is a poor approximation to

the distribution of the likelihood ratio test (and consequently, to the distribution of the score test) when there are few units and many observations per subject as in the example under discussion. Besides that, with small sample sizes, standard asymptotic tests may not yield satisfactory results. The J_n test for this example yielded a p-value equal to 0.0072. We applied both versions of the bootstrap U-test obtaining p-values equal to 0.0390 and to 0.0388, respectively for the parametric and nonparametric versions. The traditional F test provided a p-value equal to 0.04163, which would lead us to a slightly different result at the 4% significance level, for example. At the 5% significance level, both methods suggest that the variation within bulls is statistically significant. However, different p-values may lead to incorrect inference in certain situations and, in these cases, the bootstrap method seems more appropriate.

To further illustrate the advantage of the proposed test, consider a hypothetical data set generated via the following unbalanced model

$$y_{ij} = \mu + b_i + e_{ij},\ i = 1, \ldots 5,\ j = 1, \ldots, n_i$$

with $n_1 = 5, n_2 = 3, n_3 = 4, n_4 = 3$ and $n_5 = 6$. We set $\mu = 8$, $b_i \sim \{W_i - \mathbb{E}[W_i]\}/\sqrt{\text{Var}[W_i]} \times \sigma_b$, with $\sigma_b = \sqrt{0.5}$, and $e_{ij} \sim \{W_{ij} - \mathbb{E}[W_{ij}]\}/\sqrt{\text{Var}[W_{ij}]}$, where W_i and W_{ij} are *iid* random variables following St(0,1,1,4.1). The generated values are displayed in Table 4.

INSERT TABLE 4 HERE

The plot of the treatment means ($\pm$ standard errors) displayed in Figure 1 suggest heteroskedasticity, as expected, given the data were generation process.

INSERT FIGURE 1 HERE

An estimate of the mean μ is $\widehat{\mu} = \overline{y}_{\cdot\cdot} = 7.92$. Assuming normality, we obtain $\widehat{\sigma}_b^2 = 0.5961$ and $\widehat{\sigma}_e^2 = 1.0693$, so that an estimate of the intraclass correlation coefficient is $\widehat{\rho} = 0.3579$, which are values close to the true values (0.5, 1 and 1/3, respectively). The estimate based on the consistent estimator W_n is equal to 1.0241 which was the value used to implement the parametric Bootstrap test. The objective is to test whether the within-treatment variance may be dropped from the model, that is, to test (2).

For the generalized likelihood ratio test, we obtained a p-value of the 0.3575. The traditional F test provided a p-value equal to 0.1024 suggesting an inconsistent conclusion with the generating model. Here, for the J_n test, we obtained a p-value less than 10^{-7}, which may not be realistic given the small sample size. We applied both versions of the bootstrap U-test obtaining p-values of 0.0112 and 0.0335, respectively for the parametric and nonparametric versions. At the 5%

significance level, both suggest that the variation within treatment is statistically significant, as opposed to the conclusion based on the competing tests.

6 Discussion and conclusion

Although there exists an exact F-test with optimal properties for testing the significance of the between-treatments variance component in a one-way random effects model with balanced data under normality, we must rely on sub-optimal or approximate tests in unbalanced or nonnormal settings. Nobre et al. (2008) derived an asymptotic U-test that may be employed with unbalanced data and does not require a specified form for the underlying probability distributions. The authors conclude that the F-test is more affected by the lack of normality of the random effects and within-treatment errors than by imbalance. Furthermore, the U-test seems to be less sensitive to imbalance and to be more powerful than the F-test in general. Such conclusions must be viewed with caution, given the liberal nature of the U-test, specially for small sample sizes.

Sinha (2009) obtains the exact distribution of the score statistic to test the hypothesis of null variance of a random intercept in generalized linear mixed models using parametric bootstrap. Under this setup, for each pseudosample it is necessary to estimate the set of parameters and to obtain the score statistic. This requires a high computational cost given that matrix inversion, for example, may be needed. Our proposal is relatively simple, free of distribution assumptions besides presenting a very low computational cost. An extension of the U-test to more general Linear Mixed Models is proposed in Nobre et al. (2013) but its structure is slightly different from the one considered in Nobre et al. (2008) on which the bootstrap version is based. The bootstrap procedure for the U-tests in this more general situation where even under the null hypothesis, the dependent variables may not be identically distributed nor be independent is more complicated and is the object of ongoing investigation. See Davison and Hinkley (1997) and Lahiri (2003), among others, for details.

We propose bootstrap methods to obtain the empirical distribution an U-test statistic under the null hypothesis. The statistic J_n suggests the null hypothesis to be reject when high values of J_n are observed. Thus we use its easy structure to propose exact test versions using Bootstrap, both in parametric and non-parametric approach. The simulation results suggest that even for small sample sizes the test behaves well, despite a very small bias. Given that it is a test addressed specifically at small sample sizes, computational effort is not really a problem. We also evaluated the effect of misspecification of the distribution of conditional errors. We noticed that for the

parametric bootstrap, the result does not vary too much even when pseudosamples are generated from a normal distribution. The simulation codes can be obtained directly from the first author upon request.

Acknowledgements

We are grateful to Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil [grants 305336/2017-7 and 304126/2015-2] for partial financial support. We also appreciate the enlightening comments of the anonymous referees and associate editor.

References

[1] Alkharmisi M. (2000). *Asymptotic analysis of the one-way random effects models*. Phd Thesis. University of Toronto, Graduate Department of Statistics, Toronto.

[2] Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* , **65**, 367-389.

[3] Chernick, M.R. and LaBudde, R.A. (2011). *Bootstrap Methods with Application to R*. New York: John Wiley & Sons.

[4] Crainiceanu, C.M (2008). Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. In *Random Effect and Latent Variable Model Selection*. Ed: David B. Dunson. New York: Springer, Lecture Notes in Statistics # 192, 3-18.

[5] Crainiceanu, C.M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society B*, **66**, 165-185.

[6] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

[7] Demidenko, E. (2013). *Mixed models: Theory and Applications with R*, 2nd edition. New York: John Wiley & Sons.

[8] Giampaoli,V. and Singer, J.M. (2009). Generalized likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference*, **139**, 1435-1448.

[9] Greven, S., Crainiceanu, C.M., Küchenhoff, H. and Peters, A. (2008). Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. *Journal of Computational and Graphical Statistics*, **17**, 870-891.

[10] Hall, D. and Praestgaard, J.T. (2001). Order-restricted score tests for homogeneity in generalised linear and nonlinear mixed models. *Biometrika*, **88**, 739-751.

[11] Khuri, A.I., Mathew, T. and Sinha, B.K. (1998). *Statistical Tests for Mixed Linear Models*. New York: John Wiley & Sons.

[12] Kowalski, J., Pagano, M. and DeGruttola, V. (2002). A nonparametric test of gene region heterogeneity associated with phenotype. *Journal of the American Statistical Association*, **97**, 398-408.

[13] Kowalski, J. and Tu, X. M. (2007). *Modern Applied U-Statistics*. New York: John Wiley & Sons.

[14] Lee, A.J. (1990). *U-statistics: Theory and practice*. New York: Marcel Dekker.

[15] Lahiri, S.N. (2003). *Resampling Methods for Dependent Data*. New York: Springer-Verlag.

[16] Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypothesis,* 3^{rd} *Edition*. New York: Springer-Verlag.

[17] Lencina, V.B., Singer, J.M. and Stanek III, E.J. (2005). Much ado about nothing: the mixed models controversy revisited. *International Statistical Review*, **73**, 9-20.

[18] Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, **84**, 309-326.

[19] McCulloch, C.E., Searle, S.R. and Neuhaus, J.M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edition. New York: John Wiley & Sons.

[20] Nobre, J.S. (2007). Test for variance components using U-statistics. Unpublished P.h.D. thesis (in Portuguese). Departamento de Estatística, Universidade de São Paulo, Brazil.

[21] Nobre, J.S., Singer, J.M. and Silvapulle, M.J. (2008). U-tests for variance components in one-way random effects models. In *Beyond Parametrics in Interdisciplinary Research, Festschrift to P.K. Sen*. Eds: N. Balakrishnan; E. Pena; M. J. Silvapulle. California: IMS Lecture Notes-Monograph Series, Hayward, 197-210.

[22] Nobre, J.S., Singer, J.M. and Sen, P.K. (2013). Utests for variance components in linear mixed models. *TEST*, **22**, 580-605.

[23] Pinheiro, A., Sen, P.K. and Pinheiro, H.P. (2009). Decomposability of high-dimensional diversity measures: Quasi U-statistics, martingales and nonstandard asymptotics. *Journal of Multivariate Analysis*, **100**, 1645-1656.

[24] Savalli, C., Paula, G.A. and Cysneiros, F.J.A. (2006). Assesment of variance components in elliptical linear mixed models. *Statistical Modelling*, **6**, 59-76.

[25] Sen, P.K., Singer, J.M., Pedroso de Lima, A.C. (2010). *From finite sample to asymptotic methods in Statistics*. Cambridge University Press: New York.

[26] Silvapulle, M.J. and Silvapulle, P. (1995). A score test against one-sided alternatives. *Journal of the American Statistical Association*, **90**, 342-349.

[27] Silvapulle, M.J. and Sen, P. K. (2005). *Constrained Statistical Inference*. New York: John Wiley & Sons.

[28] Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *The Canadian Journal of Statistics*, **37**, 219-234.

[29] Snedecor, G.W., and Cochran, W.G. (1980). *Statistical Methods*, 7th ed. Iowa State College Press: Iowa, .

[30] Stram, D.O. and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171-1177.

[31] Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, **59**, 254-262.

[32] Zhang, D. and Lin, X. (2008). Variance Components Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and Other Related Topics. In *Random Effect and Latent Variable Model Selection*. Ed: David B. Dunson. New York: Springer, Lecture Notes in Statistics # 192, 19-36.

[33] Zhu, Z. and Fung, W.K. (2004). Variance component testing in semiparametric mixed models. *Journal of Multivariate Analysis*, **91**, 107-118.

Table 1 *Rejection rates (%) for the parametric bootstrap U-test in balanced designs with $b_i \sim \mathcal{N}(0, \sigma_b^2)$ for different distributions of the conditional errors.*

		$k=5$				$k=10$				$k=30$				$k=100$			
σ_b^2	α	$m=2$	$m=4$	$m=5$	$m=10$	$m=2$	$m=4$	$m=5$	$m=10$	$m=2$	$m=4$	$m=5$	$m=10$	$m=2$	$m=4$	$m=5$	$m=10$
								$e_{ij} \sim \mathcal{N}(0,1)$									
	1%	1.0	1.0	1.0	0.9	1.0	1.0	1.1	1.1	1.1	1.1	1.0	1.0	0.9	1.1	1.0	1.0
0	5%	5.4	5.0	5.1	5.0	5.4	5.0	5.0	5.1	5.3	5.1	5.0	4.9	5.0	5.0	5.1	5.0
	10%	10.2	10.2	10.2	10.0	10.2	10.0	10.2	10.2	10.2	10.0	10.0	9.9	10.3	9.8	10.1	10.0
	1%	2.2	7.1	10.5	30.5	3.0	13.1	20.9	56.6	7.8	38.8	56.8	95.5	25.8	90.5	98.2	100.0
0.2	5%	9.3	20.2	25.9	49.4	12.8	32.0	41.7	74.7	22.8	63.9	77.6	98.6	51.1	97.2	99.6	100.0
	10%	17.3	30.9	37.2	60.1	22.1	45.1	54.4	82.6	35.0	75.3	86.6	99.3	65.5	98.5	99.9	100.0
	1%	4.3	22.2	32.5	64.9	9.0	44.8	60.2	91.2	31.7	91.4	97.2	100.0	87.3	100.0	100.0	100.0
0.5	5%	16.2	43.6	53.5	79.1	26.9	67.6	78.6	96.6	58.3	97.2	99.4	100.0	96.3	100.0	100.0	100.0
	0%	27.7	55.7	64.6	85.1	41.2	77.8	85.6	97.9	71.8	98.7	99.8	100.0	98.6	100.0	100.0	100.0
	1%	9.0	45.9	58.3	85.1	23.2	77.3	88.3	98.8	74.1	99.8	100.0	100.0	99.9	100.0	100.0	100.0
1.0	5%	28.2	66.8	75.4	91.9	49.2	90.26	95.4	99.6	91.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10%	42.8	76.5	82.5	94.6	63.9	94.2	97.2	99.8	95.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0
								$e_{ij} \sim t_5 \times \sqrt{3/5}$									
	1%	0.7	0.7	0.7	0.8	0.8	0.8	0.9	0.9	0.8	0.9	0.9	1.0	1.1	0.9	1.0	1.0
0	5%	4.2	4.0	4.2	4.7	4.4	4.5	4.6	4.8	4.4	4.6	4.6	4.8	4.7	4.8	4.9	5.0
	10%	8.7	8.8	9.2	9.5	9.3	9.4	9.6	9.9	9.4	9.5	9.6	9.9	10.1	9.9	10.1	10.1
	1%	2.0	7.0	9.8	29.9	2.8	13.0	20.1	56.0	7.0	38.3	56.4	95.4	25.6	90.2	97.9	100.0
0.2	5%	9.0	20.1	25.4	49.2	12.1	31.7	40.7	74.4	22.4	62.9	77.2	98.3	50.6	97.1	99.6	100.0
	10%	17.2	30.8	36.7	59.9	20.9	44.2	53.7	81.5	34.2	74.0	85.9	99.0	64.6	98.2	99.8	100.0
	1%	4.1	21.3	32.2	64.0	8.6	44.0	58.5	91.0	31.2	91.1	97.1	100.0	87.0	100.0	100.0	100.0
0.5	5%	16.1	42.4	53.0	79.0	25.9	67.0	78.0	96.1	57.7	97.0	99.2	100.0	96.1	100.0	100.0	100.0
	10%	27.1	54.6	64.1	84.5	39.8	77.2	85.2	97.2	71.1	98.5	99.7	100.0	98.4	100.0	100.0	100.0
	1%	8.7	45.1	57.5	85.0	22.4	77.1	87.8	98.1	72.8	99.7	99.9	100.0	99.9	100.0	100.0	100.0
1.0	5%	28.0	66.0	75.4	91.6	49.1	89.7	95.2	99.6	90.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0
	10%	42.2	75.0	82.5	94.2	63.4	93.9	97.0	99.7	95.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
								$e_{ij} \sim (\chi^2_2 - 2)/2$									
	1%	1.2	1.2	1.1	0.9	1.2	1.1	1.1	0.9	1.2	1.1	1.1	0.9	1.1	0.1	0.9	1.0
0	5%	4.6	4.7	4.8	5.0	5.7	5.3	5.1	5.0	5.3	5.2	5.3	5.1	5.1	4.9	5.0	5.1
	10%	9.5	9.5	9.6	10.3	10.4	10.3	10.2	10.1	10.3	10.3	10.2	10.2	10.2	9.9	10.2	9.9
	1%	2.0	7.0	10.1	29.3	3.0	12.9	20.2	55.5	7.8	38.6	55.7	95.4	25.5	90.2	98.1	100.0
0.2	5%	9.1	19.9	25.1	48.6	12.1	31.4	41.2	74.3	22.7	63.2	77.5	98.5	50.9	97.1	99.6	100.0
	10%	16.8	30.3	36.8	59.3	20.9	44.3	54.3	82.2	34.9	75.0	86.1	99.1	64.5	98.4	99.8	100.0
	1%	4.2	22.0	31.2	64.5	8.9	44.6	58.1	91.1	31.3	90.9	97.1	100.0	87.1	100.0	100.0	100.0
0.5	5%	15.7	43.3	52.0	78.9	26.5	66.9	78.3	96.2	58.0	97.1	99.2	100.0	96.2	100.0	100.0	100.0
	10%	27.0	55.2	63.1	84.4	40.3	76.7	85.3	97.8	71.4	98.7	99.6	100.0	98.4	100.0	100.0	100.0
	1%	8.8	44.8	57.9	84.7	22.6	77.2	87.8	98.7	73.5	99.7	100.0	100.0	99.8	100.0	100.0	100.0
1.0	5%	27.4	65.8	75.2	91.9	48.9	89.8	94.8	99.4	90.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10%	42.1	75.0	82.4	94.3	63.3	93.7	97.0	99.7	95.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 2 *Rejection rates (%) for the nonparametric bootstrap U-test in balanced designs with $b_i \sim \mathcal{N}(0, \sigma_b^2)$ for different distributions of the conditional errors.*

		$k=5$				$k=10$				$k=30$				$k=100$			
σ_b^2	α	$m=2$	$m=4$	$m=5$	$m=10$	$m=2$	$m=4$	$m=5$	$m=10$	$m=2$	$m=4$	$m=5$	$m=10$	$m=2$	$m=4$	$m=5$	$m=10$
								$e_{ij} \sim \mathcal{N}(0,1)$									
	1%	0.5	1.0	1.1	1.1	1.0	1.0	1.1	1.1	1.1	1.2	1.2	1.1	1.3	1.0	1.1	1.1
0	5%	4.2	4.5	4.7	5.3	4.6	4.9	5.1	5.2	5.1	5.3	5.2	5.1	5.4	4.9	5.2	5.1
	10%	9.9	9.8	10.0	10.4	9.7	9.8	10.0	10.1	10.2	10.2	10.3	10.0	10.4	10.0	9.8	9.9
	1%	1.0	6.8	10.3	31.5	3.0	13.1	20.2	55.1	7.5	39.1	56.2	94.9	25.5	90.4	97.7	100.0
0.2	5%	8.2	20.5	26.2	49.8	11.6	32.3	41.0	72.8	22.8	63.7	77.4	98.4	50.9	97.1	99.5	100.0
	10%	16.9	32.2	37.3	60.5	21.3	45.1	54.0	81.0	35.0	75.1	86.2	99.2	65.4	98.4	99.7	100.0
	1%	1.8	22.0	31.4	64.6	8.0	44.6	59.5	91.0	31.4	91.0	97.1	100.0	87.1	99.9	100.0	100.0
0.5	5%	13.5	42.9	51.6	78.8	25.8	67.0	77.9	96.2	59.3	97.0	99.3	100.0	96.3	100.0	100.0	100.0
	0%	27.7	55.4	62.6	84.2	39.7	77.2	85.5	97.8	71.3	98.3	99.8	100.0	98.5	100.0	100.00	100.0
	1%	4.5	44.2	57.9	84.9	21.1	77.1	87.3	98.6	72.6	99.6	99.9	100.0	99.9	100.0	100.0	100.0
1.0	5%	25.2	65.6	75.2	91.7	48.3	90.0	94.5	99.7	90.5	99.9	100.0	100.0	100.0	100.0	100.0	100.0
	10%	40.0	76.0	82.3	94.1	62.9	93.9	96.7	99.7	95.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
								$e_{ij} \sim t_5 \times \sqrt{3/5}$									
	1%	0.4	0.9	1.1	1.2	0.7	1.1	1.1	0.9	1.3	1.2	1.3	1.1	1.2	1.0	1.2	1.2
0	5%	4.0	4.3	5.4	5.5	4.8	5.6	5.2	5.0	5.3	4.5	5.6	4.9	5.2	4.7	5.3	5.2
	10%	8.7	9.3	10.1	10.4	9.6	10.9	9.8	10.4	10.4	9.9	10.6	10.3	10.4	9.5	10.4	10.2
	1%	1.1	8.3	12.6	32.2	3.0	14.9	22.0	56.7	7.6	39.8	57.5	95.1	25.8	91.0	98.0	100.0
0.2	5%	7.8	23.0	27.4	50.9	13.0	34.0	43.3	74.4	23.7	64.2	78.5	98.9	51.5	97.3	99.7	100.0
	10%	16.7	34.4	38.6	61.4	23.1	46.3	55.2	83.0	36.5	75.6	86.1	99.6	66.9	98.8	99.9	100.0
	1%	2.5	24.2	33.1	66.2	9.4	48.1	62.6	91.1	33.5	92.1	97.2	100.0	87.6	100.0	100.0	100.0
0.5	5%	16.8	44.6	54.2	79.4	29.0	69.7	79.8	97.0	60.7	98.0	99.7	100.0	96.8	100.0	100.0	100.0
	10%	29.7	56.7	65.2	85.0	44.1	79.8	86.6	98.3	73.6	99.0	100.0	100.0	98.9	100.0	100.0	100.0
	1%	5.5	47.1	60.0	85.8	22.8	78.9	87.6	98.7	73.3	99.8	100.0	100.0	100.0	100.0	100.0	100.0
1.0	5%	26.9	67.7	76.8	92.0	50.9	90.5	94.7	99.7	90.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10%	42.3	77.4	83.3	94.4	64.3	94.1	97.0	99.8	95.3	100.0	100.0	100.0	100.0	100.0	100.0	100.0
								$e_{ij} \sim (\chi_2^2 - 2)/2$									
	1%	0.3	1.0	1.0	0.9	0.7	0.9	1.0	1.2	1.0	1.0	1.1	1.0	1.09	1.1	1.1	0.9
0	5%	4.4	4.7	4.7	4.7	4.6	5.1	4.7	4.8	4.8	5.1	4.9	4.9	4.66	5.4	5.3	4.9
	10%	9.2	9.3	9.5	9.5	9.6	10.2	9.3	9.8	10.5	10.3	9.8	9.9	10.40	9.7	10.3	9.9
	1%	1.3	9.3	13.2	34.3	4.0	16.4	23.7	59.0	8.2	41.0	58.7	95.3	26.0	90.5	98.9	100.0
0.2	5%	9.9	24.9	29.9	52.6	14.4	36.3	44.9	76.2	23.9	65.8	78.9	98.9	51.7	97.1	99.8	100.0
	10%	18.7	37.4	42.2	62.2	23.8	48.7	57.9	83.3	37.0	76.5	86.7	99.6	67.2	98.4	100.0	100.0
	1%	3.3	27.6	36.1	66.8	11.6	49.1	62.9	91.3	35.5	93.9	98.3	100.0	87.7	100.0	100.0	100.0
0.5	5%	18.7	46.2	56.3	80.0	31.2	71.0	80.2	97.0	62.3	98.7	99.9	100.0	97.2	100.00	100.0	100.0
	10%	32.5	58.9	66.1	85.8	45.5	80.4	87.3	98.8	75.8	99.4	100.0	100.0	99.0	100.00	100.0	100.0
	1%	7.6	49.4	61.3	86.3	25.6	79.7	87.9	98.8	76.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1.0	5%	30.1	69.4	77.5	92.8	52.2	92.0	95.0	99.9	92.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10%	46.5	77.9	84.0	94.9	66.9	95.3	97.2	100.0	96.1	100.0	100.0	100.00	100.0	100.0	100.0	100.0

Table 3 *Percentage of conceptions to services for sucessive samples.*

Bulls					
1	2	3	4	5	6
46	70	52	47	42	35
31	59	44	21	64	68
37		57	70	50	59
62		40	46	69	38
30		67	14	77	57
		64		81	76
		70		87	57
					29
					60

Table 4 *Hypothetical data.*

Treatment				
1	2	3	4	5
8.05	6.66	8.51	11.10	6.03
9.73	7.32	8.03	7.32	9.11
8.63	8.45	8.52	7.94	6.15
8.25		7.56		6.89
8.31				6.61
				7.05

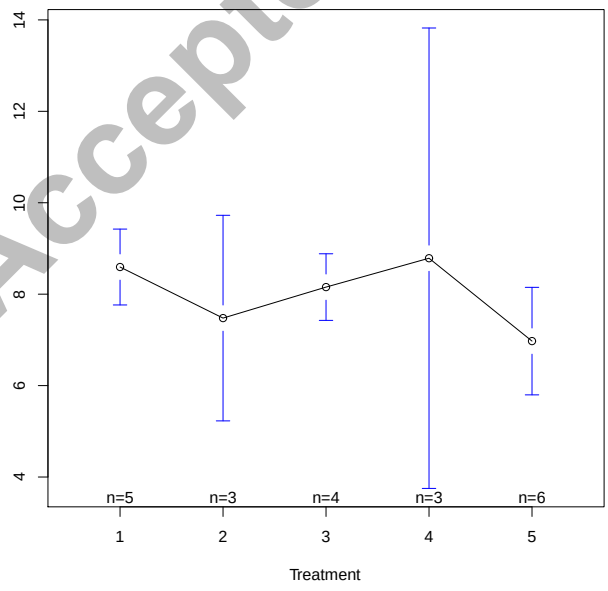

Figure 1 *Mean by treatment± standard error of the mean.*

Corresponding author. Universidade Federal do Ceará, Campus do Pici, Bloco 910, Departamento de Estatística e Matemática Aplicada, U