A COMPARISON OF VARIANCE COMPONENT TESTS

ALEXANDER F. BRUYNS[1]

An analytical evaluation is performed of four tests regarding the variance component in a one factor random effects model with a balanced design. The actual significance value and power of the asymptotic tests are computed by relating the test statistics to the *F*-distribution. In the light of these considerations a uniform arrangement of the tests is presented.

Key words: random effects model, variance component hypothesis testing, balanced design, multilevel analysis, limited parameter space, likelihood ratio test, *F* test, validity and power

1 Introduction

Multi-level analysis (Goldstein, 1987; Bryk & Raudenbush, 1992) has recently attracted the interest to the development and applications of level-structured data. The hierarchical models provide a description of the variability within and across nested levels. Primary concerns are on the estimation and hypothesis testing of variance and covariance components. As a rule multi-level analysis refers to unbalanced settings and, therefore, relies on approximate tests derived from asymptotical theories. In order to give an indication of the accuracy of these test procedures we will provide some finite sample evidence. Only in a one-way random effects model this is possible with an analytical approach. In this paper we evaluate the small sample properties of the asymptotic tests for variance components in a balanced design. The following step might be an examination of the tests for unbalanced data, which should be carried out with the use of a simulation study.

The concept of hierarchical models (Bryk and Raudenbush, 1992) is extensively used in applied research. We will present the general framework of observations within groups by means of a model at an individual level (also designated as level-1 units) are nested with groups (level-2 units). A review of four variance component tests will be made in a specific case of the linear hierarchical model, namely the one-factor random effects model in a balanced setting, presented in the next section. An outline of the available tests on the variance of the random effect will be given in section 3. Section 4 of this paper consists of a comparison of these tests on statistical grounds. Finally, we will make some remarks on this study.

---

[1] Alexander F. Bruyns is Consultant at Datastep SAS Software Consulting in Eindhoven.

2 The model and its assumptions

In order to investigate the differences between groups on the scores of their members we regard the following linear relationship at the individual level,

$$y_{ij} = \beta_j + e_{ij}, \quad (2.1)$$

between y_{ij}, the observation of i^{th} level-1 unit of group j, and the level-1 coefficient, β_j and e_{ij} ,being the level-1 error. In fact, the equation (2.1) predicts the outcome for each group with the intercept, β_j,. At level-2 we assume that each group has its own inherent β_j and decompose this effect into the true grand mean in the population denoted by μ and the random level-2 parameter, b_j:

$$\beta_j 1 = \mu + b_j .$$

Suppose, that the experiment is performed on n randomly selected level-1 units from J randomly chosen groups. Combining the aboved-mentioned formulas, we obtain

$$y_{ij} = \mu + b_j + e_{ij} \quad i = 1,\ldots,n, \qquad j = 1,\ldots J \quad (2.2)$$

in fact the conventional one-way balanced random effects model (Mason et al., 1983), which can be regarded as a modification of the corresponding fixed effects model. The groups are assumed to have been chosen randomly from a superpopulation of groups; this enables the experimenter to generalize beyond the sample being taken and to make inferential statements about the superpopulation. In order to examine the validity of these statements, however, the model requires assumptions about the observations before a test can be carried out. We assume that the random variables are independent and identically distributed as

$$b_j \sim N(0,\sigma^2_b), \qquad e_{ij} \sim N(0,\sigma^2_e),$$

where b_j and e_{ij} are mutually independent.

Important questions about a random effects model concern the unknown parameters of the population of potential groups from which our groups were selected, the mean μ or the variances σ_b^2 and σ_e^2. We focus on the absence of the population variability of level-2 effects,

$$H_0: \sigma^2_b = 0.$$

This null hypothesis holds when all unknown group effects b_j are fixed and equal.

An important statistic that gives information about this issue is the ratio of the between-groups- and the residual within group sums of squares, which in the case of a balanced design are mutually independent and defined respectively as

$$B = n \sum_{j=1}^{J} (\bar{y}_{.j} - \bar{y}_{..})^2, \quad W = \sum_{i=1}^{n} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{.j})^2 .$$

3 Tests on $H_0:\sigma_b^2=0$

Bryk and Raudenbush (1992) mentioned three large sample tests for this single parameter hypothesis: a χ^2 test, the generalized likelihood-ratio test, and a t-test based on the the ML estimator for σ_b^2 and its standard error. As is suggested by these authors, it is better to skip the last option, since under the null hypothesis the behaviour of this test results in a symmetric acceptance region for σ_b^2, which is unsatisfactory. In our balanced case we can extend the list with the usual F test, and a modified likelihood ratio test proposed to the author by Snijders. Only for the F-statistic the distribution is known, the other tests are based on an asymptotic distribution. Our list, therefore, contains four tests as potential candidates in investigating the variance of the group effect; two generalized likehood ratio tests, the usual F test, and a χ^2 approximation to that F test. A more detailed description of the four procedures is as follows.

In the case of the random effects model (2.2) we combine the likelihood function of y_{ij} conditional on b_j with the likelihood function of the group effect (see Aitkin (1989) for detials). This leads to the following profile likelihood in $\theta=\sigma_b^2/\sigma_e^2$

$$PL(\theta)=\left[\frac{W+B/(1+n\theta)}{nJ}\right]^{-nJ/2}(1+n\theta)^{-J/2}.$$

Note that $\theta \geq 0$, while the null hypothesis is $\theta=0$. This implies that the null hypothesis corresponds to a boundary region of the parameter space and the maximum likelihood estimator of θ can indeed be equal to the boundary value 0. If $(n\text{-}1)B/W > 1$, the M.L.E. of σ_b^2 and σ_e^2 are

$$\hat{\sigma}_b^2=\frac{1}{n}\left[\left(1-\frac{1}{J}\right)\frac{B}{J-1}-\frac{W}{(n-1)J}\right],$$

$$\hat{\sigma}_e^2=\frac{W}{(n-1)J}$$

and the maximum of PL(θ) is assumed for $1+n\theta=(n\text{-}1)B/W$,

$$PL(\hat{\theta})=\left[\frac{W}{(n-1)J}\right]^{-nJ/2}\left[\frac{(n-1)B}{W}\right]^{-J/2}.$$

For $(n\text{-}1)B/W \leq 1$, we obtain the following maximum likelihood estimators.

$$\hat{\sigma}_b^2=0,$$

$$\hat{\sigma}_e^2=\frac{1}{n}\left[\left(1-\frac{1}{J}\right)\frac{B}{J-1}+(n-1)\frac{W}{(n-1)J}\right].$$

In total, the likelihood ratio for the model where $\theta \geq 0$ is given by

$$\lambda=\begin{cases} 1 & for\ (n-1)B/W \leq 1 \\ \lambda_1=\dfrac{n^{nJ/2}}{(n-1)^{\frac{(n-1)J}{2}}}\left[1+\dfrac{B}{W}\right]^{-nJ/2}\left[\dfrac{B}{W}\right]^{J/2} & otherwise \end{cases}$$

According to the usual theory of the generalized likelihood ratio tests, which does not take the $\theta \geq 0$ into account, the asymptotic distribution of $-2\ln\lambda_1$ is a χ^2 with one degree of freedom, which means that we will reject the null hypothesis when $-2\ln\lambda_1$ is larger then $\chi^2_{1-2\alpha}(1)$ [Test 1]. From Self & Liang (1987) and Chernoff (1954) we know that the approximate distribution of $-2\ln\lambda$ under H_0 behaves like a 50% : 50% mixture of a χ^2 (0) (mass point at zero) and the earlier mentioned χ^2 (1). The rejection-rule for this modified LR-test is given by $2\ln\lambda > \chi^2_{1-\alpha}(1)$ [Test 2].

The next test is based on the common F statistic [Test 3]. The null hypothesis in question can be regarded as a hypothesis on differences between group means in a fixed effects model. Since under normality the sums of squares B and W are independently distributed as $(\sigma_e^2 + n\sigma_b^2)\chi^2(J\text{-}1)$ and $\sigma_e^2\chi^2(J(n\text{-}1))$, respectively, this is an exact method. Herbach (1959) shows that this test is uniformly most powerful similar and invariant in a balanced design. Moreover, in a balanced one-way classification, the exact LR test appears to be equivalent to this F-test. When σ_b^2 is equal to zero we have the exact distribution

$$F \equiv \frac{B/(J-1)}{W/(J(n-1))} \sim F(J-1, J(n-1)) \qquad (3.1)$$

However, the alternative hypothesis distributions differ between fixed and random effects models. In the case of a random factor one can derive that the F-ratio (3.1) is exactly distributed as a central F (3.1) multiplied with a constant c,

$$c \equiv \left(1 + \frac{n\sigma_b^2}{\sigma_e^2}\right).$$

Finally we employ another large sample test of $H_0{:}\sigma_b^2 = 0$, recommended by Bryk & Raudenbush, which is the χ^2 approximation to the above-mentioned F test. Under the null hypothesis, the statistic in question

$$\frac{n\sum_{j=1}^{J}(\bar{y}_{.j} - \bar{y}_{..})^2}{\hat{\sigma}_e^2} = \frac{B}{W/J(n-1)} = (J-1)F \qquad (3.2)$$

has approximately a χ^2 distribution with J-1 degrees of freedom [Test 4].

Summarizing we can test the present hypothesis by four tests, namely the generalized LR-test [1], the adjusted LR-test [2], the exact F-test [3], and the χ^2 approximation to test 3 [4]. Critera to choose between the tests are the validity, i.e., the size of the test being equal to (or not greater than) the nominal significance, and the statistical power. We prefer the tests which have a correct size (the probability of type I error); if several tests would have the correct size, we would show preference for the test with highest power. In the next section we will see that it is possible to express the power and the actual significance level of the large sample tests in terms of the exact F distribution.

4 Evaluation of the tests

From the previous section we have perceived that all test-statistics are based on the ratio of the sums of squares B and W, the main ingredients of the F test. We shall indicate now how the size α and power of the three asymptotic tests can be expressed as exact probabilities of the central F-distribution.

Since the F-test is exact, the actual significance level is equal to the nominal level. Using the definitions (3.1) and (3.2) we obtain for the χ^2 approximation to the F test [test 4] a precise expression of the type I error probability,

$$\alpha = P[\, F > \chi^2_{1-\alpha}(J-1)/(J-1)\,],$$

Similarly we express the GLR statistic $-2\ln\lambda_1$ as a function $\varphi(.)$ of the F statistic (see also Miller,1977)

$$\varphi(F) \equiv -2\ln\lambda_1 = -J\left[(n-1)\ln\left(\frac{J}{J-1}\right) + n\ln(n) + \ln(F) - n\ln\left(\frac{J(n-1)}{J-1} + F\right)\right].$$

Test 2 and 3 reject H_0 whenever $\varphi(.)$ is either larger than $\chi^2_{1-\alpha}(1)$ or $\chi^2_{1-2\alpha}(1)$ respectively. The function is strictly increasing whenever F is larger than $J/(J-1)$, which is precisely the turn-over point of the likelihood-ratio in the previous section. For this case, one can formulate an exact LR test as a F-test; instead of applying a critical value based on the large sample χ^2-distribution, one can take the critical value of the F-distribution. Using the Newton- Raphson algorithm, it is now possible to calculate which F-value corresponds to the specific critial values of asymptotic tests, leading to comparable forms of the rejection-rule of the concerning tests. As a result, probability statements of the tests can be transformed in terms of percentiles of the F-distribution with each a typical critical value. The only difference between the tests is the critical value. By comparing the critical-values of the F-distribution and those of the χ^2-distributions with a significance level for all practical purposes up to and including the 25 percent (Herbach, 1959), we can order them as follows:

$$\frac{\chi^2_{1-\alpha}(J-1)}{J-1} < F_{1-\alpha}(J-1, J(n-1)) < \varphi^{-1}\left(\chi^2_{1-2\alpha}(1)\right) < \varphi^{-1}\left(\chi^2_{1-\alpha}(1)\right) \quad (4.1)$$

It is important to remember that only the F procedure is valid under the null hypothesis and the normality assumptions. Thus, the inequality (4.1) indicates that the χ^2 approximation to F is liberal and the generalized likelihood ratio tests are conservative. The modified LR test is less conservative than the unmodified LR test. To give an illustration of the difference between the type I errors of the three limiting tests and to demonstrate their convergence properties as $n \to \infty$, we will show some examples for various values of the number of groups J and the group sizes n.

TABLE 4.1 Actual significance levels with the generalized and modified LR test and the χ^2 approximation of F for various numbers of groups J and group-sizes n at the nominal significance level of 0.01 or 0.05.

		α=-0.01			α=-0.05		
n	J	LR	mLR	$\chi^2(J\text{-}1)$	LR	mLR	$\chi^2(J\text{-}1)$
5	5	.00199	.00409	.03069	.01076	.02265	.08689
5	10	.00249	.00514	.02747	.01352	.02838	.08347
5	20	.00296	.00610	.02490	.01596	.03325	.08017
5	50	.00352	.00721	.02257	.01864	.03840	.07684
10	5	.00154	.00325	.01823	.00886	.01916	.06643
10	10	.00212	.00444	.01703	.01195	.02551	.06499
10	20	.00267	.00556	.01605	.01473	.03102	.06356
10	50	.00332	.00683	.01516	.01779	.03687	.06209
20	5	.00136	.00291	.01366	.00806	.01767	.05776
20	10	.00196	.00415	.01315	.01127	.02425	.05710
20	20	.00254	.00513	.01234	.01418	.03001	.05643
20	50	.00323	.00666	.01273	.01740	.03617	.05575
50	5	.00126	.00272	.01137	.00762	.01685	.05300
50	10	.00187	.00398	.01118	.01088	.02353	.05275
50	20	.00247	.00518	.01103	.01387	.02943	.05249
50	50	.00318	.00656	.01089	.01718	.03577	.05223

With respect to the probability type I error, table 4.1 confirms inequality (4.1): the likelihood-ratio tests are conservative and the χ^2 approximation of F is liberal. The size of the modified LR is about twice as large as that of the generalized LR, but continues to remain smaller than the nominal significance value. For large group sizes the significance level of the $\chi^2(J\text{-}1)$ test is closer to the nominal α of the exact F statistic than the LR tests, especially for small group sizes; The likelihood-ratio tests show a relatively strong countereffect when the group sizes n increase in comparison with a rise in J. The net effect is however upward if the number of groups J increases with the same proportion as n. With the $\chi^2(J\text{-}1)$ test these changes of n and J work in the same direction, but are again more substantial for the group - size. Similar conclusions can be drawn for other low tail probabilities of α.

In section 3 we saw that the probability statements of the four test-statistics under the null, as well as under the alternative hypothesis, depend on the exact distribution of F statistic. For the alternative hypothesis we have the same central F-distribution, but now

modified by a positive constant c. This implies maintenance of the order in inequality (4.1), as all critical values are divided by the same value of c: the most conservative test is the least powerful and vice versa. For instance, the expression of the power of the common F-test becomes

$$\pi = P[\, F > F_{1-\alpha}(I-1, I(n-1))/c \,]$$

Notice that the power can be derived from the ordinary cumulative F-table, instead of the more complex charts of the noncentralized F in the fixed effects model.

In order to assess the power, it is necessary to specify some numerical values of the ratio θ. We will vary the population variances under the condition that their sum is equal to one. Due to the simple division of the constant c, general reflections on table 4.1 still maintain: the more observations or the higher the nominal α one is considering, the smaller the probability of type II error and thus, the larger the power. Once again, the group size exerts a greater influence than the number of groups J. To give an indication of the difference, we will show some results.

TABLE 4.2 The power of the four tests for several numbers of groups J and group sizes n if α equals 0.01 under the condition that the sum of the variance equals one.

σ_b^2	LR	mLR	F	$\chi^2(J\text{-}1)$	LR	mLR	F	$\chi^2(J\text{-}1)$
	$J=10$ $n=5$				$J=5$ $n=10$			
0.1	.03505	.05687	.08767	.16495	.05800	.08513	.14858	.19806
0.2	.15309	.20975	.27588	.40530	.22110	.27641	.37941	.44502
0.3	.36214	.44023	.51912	.64778	.42445	.48627	.58756	.64505
0.4	.59889	.66982	.73380	.82456	.60923	.66214	.74201	.78408
0.5	.79391	.84030	.87849	.92702	.75443	.79293	.84782	.87526
0.6	.91697	.93908	.95597	.97560	.85907	.88350	.91690	.93297
0.7	.97609	.98329	.98846	.99403	.92911	.94233	.95983	.96801
0.8	.99609	.99738	.99826	.99915	.97181	.97737	.98454	.98781
0.9	.99983	.99989	.99993	.99997	.99368	.99498	.99663	.99736
	$J=20$ $n=5$				$J=5$ $n=20$			
0.1	.08391	.12515	.16306	.25995	.20831	.25919	.36436	.39587
0.2	.37895	.46250	.52468	.64665	.52454	.57896	.67268	.69715
0.3	.72446	.78689	.82664	.89128	.73209	.77023	.83082	.84567
0.4	.92014	.94429	.95792	.97719	.85222	.87579	.91155	.92001
0.5	.98496	.99040	.99319	.99673	.92081	.93435	.95429	.95890
0.6	.99828	.99898	.99932	.99970	.95988	.96706	.97743	.97979
0.7	.99990	.99995	.99996	.99999	.98173	.98511	.98992	.99101
0.8	1.0000	1.0000	1.0000	1.0000	.99330	.99457	.99636	.99676
0.9	1.0000	1.0000	1.0000	1.0000	.99860	.99887	.99925	.99933

The table illustrates that power values tend to unity when we consider more observations. If we compare samples with the same total of measurements, but with a different values of n and J, then at a certain point the power turns to a higher value. Therefore just as with α, the group size n shows a greater impact in comparison to the amount of groups.

5 Final remarks and suggestions for further research

This paper has discussed the evaluation of four tests for a variance component. Only in the case of perfectly balanced designs the distributions of the F statistic under the null as well as under the alternative hypothesis are exact. This enabled us to determine the exact validity and power of related asymptotic test-procedures. An order between these tests was given and illustrations were shown to reflect the differences between these tests and the most appropriate test. The χ^2 (J-1) test is liberal, the modified LR is conservative, and the GLR is very conservative. In the overall assessement, the common F-test gives the best balance, as it represents the nominal α even though the power is lower than its χ^2 approximation. The modified LR test is to be preferred to the unmodified LR test.

For more complicated but useful designs, we have to generalize beyond the scope of the model offered in this article. The model can be extended if we include some covariates, either level-1 predictors or group-characteristics with fixed effects which predict the group means This additon will affect the definitions of the sums of squares and the degrees of freedom of the test-statistics, but in total the conception of the tests remains the same (Bryk & Raudenbush, 1992). With more random factors included in the model, the F tests remains the most attractive but shows no link with the limiting likelihood ratio tests due to the more complete boundaries to the paramater space (Herbach, 1959 and Gautschi, 1959). The relationship with the χ^2 approximation to the F test in models with multiple random factors is a topic for investigation. The proposed computations have to be extended to these situations. This large sample test is, however, also practical in the more frequent unbalanced cases. The likelihood ratio test is also suitable for the multiparameter hypothesis tests, for which the modified LR test is the most appropriate. For the unbalanced classifications improved estimators of the effects and simulation studies in these settings are needed to supplement the guidelines to the real significance value and power differences in this article. The paper already gives an indication of the suitablity of the large sample distribution in the unbalanced designs.

Acknowledgement The author would like to thank Prof. Dr. T.A.B. Snijders of the University of Groningen for his valuable advice and suggestions. His insight into the problem led to the final result of this article.

References

Aitkin, M (1989). *Profile predictive likelihood for random effects in the two-level model.*in: *Multi-level Analysis of Educational Data.* R.D. Bock.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Statistics*, **25**,573-578.

Gautschi, W. (1959) Some remarks on Herbach's paper, 'Optimum nature of the F-test for Model II in the balanced case.'. *Annals of Mathematical Statistics*, **30**, 960-963.

Goldstein, H. (1986). *Multilevel Models in Educational and Social Research.* London: Oxford University Press.

Herbach, L.H. (1959) Properties of model II-type analysis of variance tests, A: optimum nature of the F-test for model II in the balanced case. *Annals of Mathematical Statistics*, **30**, 939-959.

Miller, J.J. (1977). Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance. *The Annals of Statistics*, **5**, 746-762.

Bryk, A.S. & Raudenbush, S.W. (1992).*Hierarchical Linear Models.*Newbury Park, CA: Sage Publications.

Self, S.G. & Liang, K. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. *Journal of the American Statistical Association*, **87**, 605-610.

Ontvangen: 7-6-98
Geaccepteerd: 6-7-99