ON THE EXACT AND NEAR-EXACT DISTRIBUTIONS OF STATISTICS USED IN GENERALIZED F TESTS

Carlos A. Coelho & João T. Mexia
Mathematics Department, Faculty of Sciences and Technology,
The New University of Lisbon, 2829-516 Caparica, Portugal

SYNOPTIC ABSTRACT

Statistics that are the ratio of two independent linear combinations of independent chi-squared random variables are used to test hypotheses on parameters in mixed and random-effects models and to test hypotheses on the parameters in the joint analysis model of several experiments. We will call these statistics generalized F statistics and the associated test a generalized F test. In this paper we first obtain the exact distribution of any statistic that is the ratio of two independent linear combinations of independent Gamma distributed random variables. Based on this distribution we then obtain asymptotic and near-exact distributions for such statistics. Then, the exact, asymptotic and near-exact distributions of generalized F statistics are readily derived, under both the null andthe alternative hypotheses. Given that the exact distributions are infinite mixtures, they are not much adequate for practical purposes and thus the development of near-exact distributions is a desirable goal. Some examples of application are shown.

Key words and phrases: Generalized F tests, generalized F statistics, asymptotic distributions, near-exact distributions.

1. INTRODUCTION

Generalized F tests were introduced by Michalski & Zmyślony (1996, 1999) to test for variance compnents. The statistics used in these tests are the ratio of two independent linear combinations of independent chi-square distributed random variables. We will call these statistics, generalized F statistics.

Expressions for the exact distribution of generalized F statistics in the central case were obtained by Fonseca *et al.* (2002) for the cases where the chi-squared random variables in the numerator or in the denominator have even numbers of degrees of freedom. This result was extended by Nunes & Mexia (2004) to the non-central case.

In this paper we first obtain the exact distribution of statistics that are the ratio of two independent linear combinations of independent Gamma distributed random variables, based on the result from Moschopoulos (1985) on the distribution of sums of independent Gamma random variables. Then from this exact distributon obtained for the ratio of two independent linear combinations of independent Gamma distributed random variables we easily derive, as a particular case, the exact distribution of generalized F statistics, as well as asymptotic and near-exact distributions for such statistics.

2. THE EXACT DISTRIBUTION OF THE RATIO OF TWO INDEPENDENT LINEAR COMBINATIONS OF INDEPENDENT GAMMA RANDOM VARIABLES

2.1 *The exact distribution of a linear combination of independent Gamma random variables* (Moschopoulos, 1985)

In this subsection we settle some notation and restate a result from Moschopoulos (1985), using only a more convenient notation for our objectives.

We will say that the random variable X has a Gamma distribution with shape parameter $r > 0$ and rate parameter $\lambda > 0$ (we will call this parameter, 'rate' parameter, given its relation, for integer r, with the rate of Poisson process) if the pdf (probability density function) of X is given by

$$f_X(x) \;=\; \frac{\lambda^r}{\Gamma(r)}\, e^{-\lambda x}\, x^{r-1} \qquad (x > 0)\,.$$

We will denote this fact by

$$X \sim \Gamma(r, \lambda)\,.$$

The cdf (cumulative distribution function) of X is then given by

$$F_X(x) = P(X \leq x) = \frac{\Gamma(r, \lambda x)}{\Gamma(r)}\,,$$

where

$$\Gamma(r, \lambda x) = \int_0^\infty \lambda^r\, e^{-\lambda u}\, u^{r-1}\, du$$

is the incomplete Gamma function. The c.f. (characteristic function) of X is then

$$\Phi_X(t) = \lambda^r (\lambda - it)^{-r}\,.$$

Let

$$X_i \sim \Gamma(r_i, \lambda_i)\,, \qquad i = 1, \ldots, n \tag{1}$$

be n independent distributed random variables. We want expressions for the exact pdf, cdf and c.f. of the random variable

$$W = \sum_{i=1}^{n} a_i X_i$$

where $a_i > 0$, for all $i \in \{1, \ldots, n\}$.

To obtain the expression for the pdf of W we only have to note that from (1),

$$a_i X_i \sim \Gamma(r_i, \mu_i)$$

with

$$\mu_i = \lambda_i / a_i > 0 \tag{2}$$

and then use directly the result in Moschopoulos (1985), which, as the author aknowledges at the end of his paper, is readiy applicable to the distribution of linear combinations of independent Gamma random variables with positive coefficients. This way we obtain the pdf of W as

$$\begin{aligned} f_W(w) &= \sum_{j=0}^{\infty} w_j \frac{\mu^{r+j}}{\Gamma(r+j)}\, e^{-\mu w}\, w^{r+j-1} \\ &= C \sum_{j=0}^{\infty} \delta_j \frac{\mu^{r+j}}{\Gamma(r+j)}\, e^{-\mu w}\, w^{r+j-1} \end{aligned}$$

where

$$\mu = \max_{1\le i\le n} \mu_i \ , \qquad r = \sum_{i=1}^{n} r_i$$

and

$$w_j \ = \ C\,\delta_j$$

with

$$C \ = \ \prod_{i=1}^{n} \left(\frac{\mu_i}{\mu}\right)^{r_i} \ , \qquad \delta_j = \frac{1}{j}\sum_{k=1}^{j} \delta_{j-k}\, k\, \gamma_k \ , \quad (j = 1, 2, \ldots)$$

$$\delta_0 = 1$$

where

$$\gamma_k \ = \ \frac{1}{k}\sum_{i=1}^{n} r_i \left(1 - \frac{\mu_i}{\mu}\right)^{k} \qquad k = 1, 2, \ldots \ ,$$

with μ_i given by (2).

That this distribution is an infinite mixture of $\Gamma(r + j, \mu)$ distributions $(j = 0, 1, \ldots)$ with weights w_j is not explicitly aknowledged by Moschopoulos (1985), but indeed $\sum_{j=0}^{\infty} w_j = 1$, clearly with $w_j > 0$ $(j = 0, 1, \ldots)$.

Once acknowledged the mixture structure of the distribution of W, the cdf and the c.f. of W are readily obtained respectively as

$$F_W(w) \ = \ \sum_{j=0}^{\infty} w_j \, \frac{\Gamma(r + j, \mu w)}{\Gamma(r + j)} \ = \ C \sum_{j=0}^{\infty} \delta_j \, \frac{\Gamma(r + j, \mu w)}{\Gamma(r + j)}$$

and

$$\Phi_W(t) \ = \ \sum_{j=0}^{\infty} w_j \, \mu^{r+j} \, (\mu - it)^{-r-j} \ = \ C \sum_{j=0}^{\infty} \delta_j \, \mu^{r+j} \, (\mu - it)^{-r-j} \ .$$

2.2 *The Gamma-ratio distribution*

The Gamma-ratio distribution is the distribution of the random variable

$$Y = X_1 / X_2$$

where

$$X_1 \sim \Gamma(r_1, \lambda_1) \qquad \text{and} \qquad X_2 \sim \Gamma(r_2, \lambda_2)$$

are two independent random variables.

The pdf of Y is

$$f_Y(y) \ = \ \frac{k^{r_1}}{B(r_1, r_2)} \, (1 + ky)^{-r_1 - r_2} \, y^{r_1 - 1} \qquad (y > 0)$$

where

$$k = \lambda_1/\lambda_2\,.$$

The cdf of Y is

$$F_Y(y) \;=\; \frac{k^{r_1}}{B(r_1,r_2)}\,\frac{y^{r_1}}{r_1}\,{}_2F_1(r_1+r_2,r_1;r_1+1;-ky)$$

and the c.f. of Y is

$$\Phi_Y(t) \;=\; E\left(e^{itY}\right) \;=\; \frac{\Gamma(r_1+r_2)}{\Gamma(r_2)}\,\Psi\left(r_1,1-r_2;-\frac{it}{k}\right)$$

where

$$\begin{aligned} {}_2F_1(a,b;c;z) \;&=\; \sum_{i=0}^{\infty}\frac{\Gamma(a+i)}{\Gamma(a)}\,\frac{\Gamma(b+i)}{\Gamma(b)}\,\frac{\Gamma(c)}{\Gamma(c+i)}\,\frac{z^i}{i!} \\ &=\; \frac{\Gamma(c)}{\Gamma(b)\,\Gamma(c-b)}\int_0^1 x^{b-1}(1-x)^{c-b-1}(1-zx)^{-a}\,dx \end{aligned}$$

is the Gauss hypergeometric function, and

$$\Psi(a,b;z) \;=\; \frac{1}{\Gamma(a)}\int_0^\infty e^{-zt}\,t^{a-1}\,(1+t)^{b-a-1}\,dt\,,\quad (z\in\mathbb{C})$$

is the Tricomi hypergeometric function.

The h-th moment of Y is given by

$$E\left(Y^h\right) \;=\; k^{-h}\,\frac{\Gamma(r_1+h)}{\Gamma(r_1)}\,\frac{\Gamma(r_2-h)}{\Gamma(r_2)} \qquad (-r_1<h<r_2)\,.$$

2.3 *The exact distribution of the ratio of two independent linear combinations of independent Gamma random variables and of the generalized F statistic*

Let

$$Z = W_1/W_2$$

where

$$W_1=\sum_{i=1}^{n_1}a_iX_i \qquad \text{and} \qquad W_2=\sum_{i=1}^{n_2}b_iX_i^*$$

are two independent random variables, with

$$X_i\sim\Gamma(s_i,\lambda_i)\quad i=1,\ldots,n_1 \qquad \text{and} \qquad X_i^*\sim\Gamma(s_i^*,\lambda_i^*)\quad i=1,\ldots,n_2$$

being independent random variables.

Then, given the mixture structure of the distributions of W_1 and W_2 described in subsection 2.1, the exact distribution of Z may be easily obtained as a double mixture of Gamma-ratio distributions with pdf

$$\begin{aligned} f_Z(z) &= \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} w_j\, w_l^*\, \frac{k^{r_1+j}}{B(r_1+j, r_2+l)}\, (1+kz)^{-r_1-r_2-j-l}\, z^{r_1+j-1} \\ &= C_1 C_2 \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} \delta_j\, \delta_l^*\, \frac{k^{r_1+j}}{B(r_1+j, r_2+l)}\, (1+kz)^{-r_1-r_2-j-l}\, z^{r_1+j-1} \end{aligned} \tag{3}$$

where

$$k = \frac{\mu_1}{\mu_2}\,, \qquad r_1 = \sum_{i=1}^{n_1} s_i\,, \qquad r_2 = \sum_{i=1}^{n_2} s_i^*\,,$$
$$\mu_1 = \max_{1\le i\le n_1} \frac{\lambda_i}{a_i}\,, \qquad \mu_2 = \max_{1\le i\le n_2} \frac{\lambda_i^*}{b_i}\,,$$
$$w_j = C_1\delta_j\,, \qquad w_l^* = C_2\delta_l^*\,,$$

with

$$C_1 = \prod_{i=1}^{n_1}\left(\frac{\lambda_i}{a_i\,\mu_1}\right)^{s_i}\,, \qquad C_2 = \prod_{i=1}^{n_2}\left(\frac{\lambda_i^*}{b_i\,\mu_2}\right)^{s_i^*}\,,$$

$$\delta_j = \frac{1}{j}\sum_{k=1}^{j}\delta_{j-k}\,k\,\gamma_k\,, \qquad \delta_l^* = \frac{1}{l}\sum_{k=1}^{l}\delta_{l-k}^*\,l\,\gamma_k^* \quad \text{with} \quad \delta_0 = \delta_0^* = 1\,,$$

where

$$\gamma_k = \frac{1}{k}\sum_{i=1}^{n_1} s_i\left(1-\frac{\lambda_i}{a_i\,\mu_1}\right)^k\,, \qquad \gamma_k^* = \frac{1}{k}\sum_{i=1}^{n_2} s_i^*\left(1-\frac{\lambda_i^*}{b_i\,\mu_2}\right)^k\,.$$

We should note that in (3), $\sum_{j=0}^{\infty}\sum_{l=0}^{\infty} w_j w_l^* = 1$. The cdf of Z is given by

$$\begin{aligned} F_Z(z) &= \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} w_j\, w_l^*\, \frac{k^{r_1+j}}{B(r_1+j, r_2+l)}\, \frac{z^{r_1+j}}{r_1+j} \\ &\qquad {}_2F_1(r_1+r_2+j+l, r_1+j; r_1+1+j; -kz) \\ &= C_1 C_2 \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} \delta_j\, \delta_l^*\, \frac{k^{r_1+j}}{B(r_1+j, r_2+l)}\, \frac{z^{r_1+j}}{r_1+j} \\ &\qquad {}_2F_1(r_1+r_2+j+l, r_1+j; r_1+1+j; -kz) \end{aligned}$$

and the c.f. of Z is

$$\begin{aligned} \Phi_Z(t) &= \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} w_j\, w_l^*\, \frac{\Gamma(r_1+r_2+j+l)}{\Gamma(r_2+l)}\, \Psi\left(r_1+j, 1-r_2-l; -\frac{it}{k}\right) \\ &= C_1 C_2 \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} \delta_j\, \delta_l^*\, \frac{\Gamma(r_1+r_2+j+l)}{\Gamma(r_2+l)}\, \Psi\left(r_1+j, 1-r_2-l; -\frac{it}{k}\right). \end{aligned}$$

From this exact distribution for Z we may even easily derive the exact moments of Z, using the mixture structure, as

$$E\left(Z^h\right) = \sum_{j=0}^{\infty}\sum_{l=0}^{\infty} w_j\, w_l^*\, k^{-h}\, \frac{\Gamma(r_1+j+h)}{\Gamma(r_1+j)}\, \frac{\Gamma(r_2+l-h)}{\Gamma(r_2+l)} \qquad (-r_1-j < h < r_2+l)\,. \tag{10}$$

For the generalized F statistic, we have

$$X_i \sim \chi^2_{\nu_i} \equiv \Gamma\left(\frac{\nu_i}{2}, \frac{1}{2}\right) \qquad (i = 1, \ldots, n_1) \tag{11}$$

and

$$X_i^* \sim \chi^2_{\eta_i} \equiv \Gamma\left(\frac{\eta_i}{2}, \frac{1}{2}\right) \qquad (i = 1, \ldots, n_2) \tag{12}$$

so that the exact pdf, cdf and c.f. of

$$F^* = \frac{\sum\limits_{i=1}^{n_1} a_i X_i}{\sum\limits_{i=1}^{n_2} b_i X_i^*}$$

with X_i and X_i^* given by (11) and (12) are given respectively by (3), (4) and (5), with

$$r_1 = \frac{1}{2}\sum_{i=1}^{n_1}\nu_i\ , \qquad r_2 = \frac{1}{2}\sum_{i=1}^{n_2}\eta_i\ ,$$

$$k = \frac{\mu_1}{\mu_2} \quad \text{with} \quad \mu_1 = \max_{1\le i\le n_1}\frac{1}{2a_i}\ , \quad \mu_2 = \max 1 \le i \le n_2 \frac{1}{2b_i}$$

and all other parameters defined in a similar way with λ_i and λ_i^* replaced by $1/2$, s_i replaced by $\nu_i/2$ and s_i^* replaced by $\eta_i/2$.

3. NEAR-EXACT DISTRIBUTIONS

Asymptotic distributions for both Z and F^* may be obtained by simple truncation of the series in (3), (4) or (5). These distributions will be asymptotic in the sense that, as the number of terms in the double summation is allowed to grow indefinitely, the asymptotic distributions will tend to the exact distribution. However, given the weights in these asymptotic distributions do not add up to 1, the computation of large quantiles from these distributions is not accurate. This difficulty may be overcome and the proximity of the asymptotic distributions to the exact distribution improved by rescaling the

weights w_j and w_l^* in the double summation to have them adding up to 1, that is, by replacing w_j by

$$w_j^* = \frac{w_j}{\theta_1} \qquad \text{where} \qquad \theta_1 = \sum_{j=0}^{m_1} w_j \,, \qquad j = 0, \ldots, m_1 \,,$$

considering that we have truncated the summation over j to $m_1 + 1$ terms, and by replacing w_l^* by

$$w_l^{**} = \frac{w_l^*}{\theta_2} \qquad \text{where} \qquad \theta_2 = \sum_{l=0}^{m_2} w_l^* \,, \qquad l = 0, \ldots, m_2 \,,$$

considering that we have truncated the summation over l to $m_2 + 1$ terms.

We should note that the asymptotic distributions obtained by rescaling the weights are truly asymptotic in the sense that they converge to the exact distribution not only when the number of terms in the summations increases, but also, and more important, when the values of the shape parameters s_i and s_i^* in (3) increase. This feature, that may be better analyzed in the next section, is really much important, being the one that makes of these distributions really asymptotic distributions, since the value of the parameters s_i and s_i^* are really in many applications related with sample sizes, namely when we deal with distributions of statistics, and namely when we consider the distribution of generalized F statistics, situation in which, as we will see in section 5, the values of these shape parameters are directly related with the number of degrees of freedom of factors involved in the model.

Although the rescaling of the weights improves much the closeness of the asymtotic distributions, as we will see in the next section, we may still improve the closeness to the exact distribution by building near-exact distributions, which, as it will also be seen in section 4, will be much closer to the exact distribution than the asymptotic distributions, mainly when we consider a rather small number of terms in the summations.

Near-exact distributions for both Z and F^* may then be also derived from truncations of the c.f. of these random variables. Since the distribution of F^* is just a particular case of the distribution of Z, we will only address this latter one, being then the near-exact distributions for F^* easily derived using the replacements at the end of the previous subsection.

We may wite the c.f. of Z as

$$\Phi_Z(t) = \sum_{j=0}^{m_1}\sum_{l=0}^{m_2} w_j\, w_l^* \,\frac{\Gamma(r_1+r_2+j+l)}{\Gamma(r_2+l)}\,\Psi\left(r_1+j, 1-r_2-l; -\frac{it}{k}\right) + R_n(t) \tag{12}$$

where

$$\begin{aligned} R_n(t) &= \Phi_Z(t) - \Phi_n(t) \\ &= \sum_{j=0}^{n}\sum_{l=m_2+1}^{\infty} w_j\, w_l^* \,\frac{\Gamma(r_1+r_2+j+l)}{\Gamma(r_2+l)}\,\Psi\left(r_1+j, 1-r_2-l; -\frac{it}{k}\right) \\ &\quad + \sum_{l=0}^{n}\sum_{j=m_1+1}^{\infty} w_j\, w_l^* \,\frac{\Gamma(r_1+r_2+j+l)}{\Gamma(r_2+l)}\,\Psi\left(r_1+j, 1-r_2-l; -\frac{it}{k}\right) \\ &\quad + \sum_{j=m_1+1}^{\infty}\sum_{l=m_2+1}^{\infty} w_j\, w_l^* \,\frac{\Gamma(r_1+r_2+j+l)}{\Gamma(r_2+l)}\,\Psi\left(r_1+j, 1-r_2-l; -\frac{it}{k}\right). \end{aligned}$$

Our aim is to obtain near-exact approximations to the distribution of Z by using an adequate number of terms of the exact c.f. of Z, that is, adequate values for m_1 and m_2 in (12), and to replace $R_n(t)$ by

$$\phi_1(t) = \theta\,\lambda^r(\lambda - it)^{-r}\,, \tag{13}$$

that is the c.f. of a Gamma distribution with shape parameter r and rate parameter λ afected by an 'external' weight

$$\begin{aligned} \theta &= 1 - \sum_{j=0}^{m_1}\sum_{l=0}^{m_2} w_j\, w_l^* \;=\; 1 - C_1C_2\sum_{j=0}^{n}\sum_{l=0}^{n}\delta_j\,\delta_l^* \\ &= \sum_{j=0}^{n}\sum_{l=m_2+1}^{\infty} w_j\, w_l^* + \sum_{l=0}^{n}\sum_{j=m_1+1}^{\infty} w_j\, w_l^* + \sum_{j=m_1+1}^{\infty}\sum_{l=m_2+1}^{\infty} w_j\, w_l^*\,. \end{aligned}$$

The approximations will be done in such a way that the near-exact distributions will have either the first two moments equal to the exact ones, by requiring that

$$\left.\frac{d^h}{dt^h}\,R_n(t)\right|_{t=0} \;=\; \left.\frac{d^h}{dt^h}\,\phi_1(t)\right|_{t=0} \qquad h = 1, 2 \tag{14}$$

what will imply that $\Phi_Z(t)$, the exact c.f. of Z, and

$$\Phi_1(t) = \Phi_n(t) + \phi_1(t)$$

the near-exact c.f. of Z will have the same first two derivatives with respect to t, at $t = 0$, equal.

This way, the near-exact distribution obtained for Z is a finite mixture with $n + 1$ components, the first n of which are Gamma-ratio distributions and the remaining one is a Gamma distribution.

According to the objectives settled in obtaining this near-exact distribution, the parameters r and λ in (13) are given by

$$r = \frac{\mu_1^{*2}}{\mu_2^* - \mu_1^{*2}} \qquad \text{and} \qquad \lambda = \frac{\mu_1^*}{\mu_2^* - \mu_1^{*2}}$$

with

$$\mu_1^* = \frac{\mu_1}{i\theta} \qquad \text{and} \qquad \mu_2^* = -\frac{\mu_2}{\theta}$$

where μ_1 and μ_2 are the two first derivetives of $R_n(t)$ with respect to t, at $t = 0$, what is equivalent to solve the system of two equations resulting from (14) above, for r and λ.

The asymptotic behaviour of both the asymptotic and near-exact distributions have their expression in the fact that as the shape parameters s_i and s_i^* in (3) increase their values, the closeness of those distributions to the exact distribution improves. We should however note that this is only true for the asymptotic distributions with rescaled weights, since the ones obtained by simple truncation do not have this behaviour. We should also note that, as it may also be analyzed in more detail in the next section, the near-exact distributions also have a much accentuted asymptitc behaviour than the simple asymptotic distributions.

Although it seems that we might have considered more elaborate approximations to $R_n(t)$ when building the near-exact distributions, by equating a larger number of moments and likely getting better approximations, with good candidates for the replacement of $R_n(t)$ being either a mixture of two Gamma distributions or a Gamma-ratio distribution, the values obtained for the parameters by equating derivatives of the corresponding characteristic function and of $R_n(t)$ would not be acceptable. Moreover, as we will see in the numerical studies section, equating the two first moments leads to high performance approximations.

4. AN APPLICATION

We now apply our results to the tests for variance components in random effects models with balanced cross-nesting. Let there be L groups of $u_1, \ldots, u_L$ factors. The first factors in the groups will have $a_l(1)$ $(l = 1, \ldots, L)$ levels and, if $u_l > 1$, each level of the h-th $(h = 1, \ldots, u_l - 1)$ factor nests $a_l(h+1)$ levels of the following factor. We take $c_l(0) = 1$, while $c_l(h) = \prod_{k=1}^{h} a_l(k)$ will be the number of level combinations for the first h factors in the l-th, each of them nesting $b_l(h) = c_j(u_l)/c_l(h)$ $(h = 0, \ldots, u_l)$, level combinations of the remaining factors $(l = 1, \ldots, L)$.

Since inside each nesting group factors do not interact, factor effects and interactions will correspond to sets of at most one factor per group. These sets of factors correspond to the vectors $\underline{h}$ with components $h_l = 0, \ldots, u_l$, $l = 1, \ldots, L$. When $h_l = 0$ no factor is taken from the l-th group, otherwise h_l will be the factor index. To the vector $\underline{h}$ corresponds $c(\underline{h}) = \prod_{l=1}^{L} c_l(h_l)$ level combinations, each one nesting $b(\underline{h}) = r \prod_{l=1}^{L} b_l(h_l)$ observations if we take r replicates.

Representing by $\otimes$ the Kronecker matrix product the model may be written, see Fonseca *et al.* (2003a), as

$$\underline{y} = \sum_{\underline{h} \in \Gamma} X(\underline{h}) \underline{\beta}(\underline{h}) + \underline{e}$$

where Γ is the set of vectors $\underline{h}$,

$$X(\underline{h}) = \bigotimes_{l=1}^{L} X_l(h_l)\,; \qquad h \in \Gamma\,,$$

with

$$X_l(0) = \underline{1}_{b_l(0)}\,; \quad l = 1, \ldots, L$$

and

$$X_l(h) = I_{c_l(h)} \otimes \underline{1}_{b_l(h)}\,, \quad h = 1, \ldots, u_l;\ l = 1, \ldots, L$$

while, with μ the general mean, $\underline{\beta}(\underline{0}) = \mu$, the remaining $\beta(\underline{h})$, and $\underline{e}$ are assumed to be Normal, independent, with null mean vectors and variance-covariance matrices $\sigma^2(\underline{h}) I_{c(\underline{h})}$ $(\underline{h} \neq 0)$ and $\sigma^2 I_n$, where $n = r \prod_{l=1}^{L} c_l(u_l)$ is the number of observations.

Following Michalski & Zmyślony (1996) we take as test statistic for

$$H_0(\underline{h}):\ \sigma^2(\underline{h})=0,\ \underline{h}\neq\underline{0}$$

the ratio of the positive by the negative part of a quadratic unbiased estimator of $\sigma^2(\underline{h})$, $\underline{h}\neq\underline{0}$. Such an estimator is, see Fonseca *et al.* (2003a),

$$\tilde{\sigma}^2(\underline{h})=\frac{1}{b(\underline{h})}\sum_{\underline{k}\in\Theta(\underline{h})}(-1)^{m(\underline{h},\underline{k})}\frac{s(\underline{k})}{g(\underline{k})}$$

where $\Theta(\underline{h})$ is the set of vectors $\underline{k}$ with components k_l such that $h_l\leq k_l\leq \min(h_l+1,u_l)$, $(l=1,\ldots,L)$, while $m(\underline{h},\underline{k})$ is the number of the components of $\underline{h}$ which are smaller than the corresponding components of $\underline{k}$. Moreover, $g(\underline{k})=\prod_{l=1}^{L}g_l(k_l)$ with $g_l(0)=1$ and $g_l(k)=c_l(k)-c_l(k-1)$, $k=1,\ldots,u_l; l=1,\ldots,L$ and

$$s(\underline{k})=||A(\underline{k})\underline{y}||^2\,;\quad \underline{k}\in\Gamma\,,$$

with

$$A(\underline{k})=\bigotimes_{l=1}^{L}A_l(k_l)\,;\quad \underline{k}\in\Gamma\,,$$

where

$$\begin{cases} A_l(0)=\dfrac{1}{\sqrt{b_l(0)}}\underline{1}_{b_l(0)}^T\,;\quad l=1,\ldots,L \\ \\ A_l(k)=I_{c_l(k-1)}\otimes K_{a_l(k)}\otimes\left(\dfrac{1}{\sqrt{b_l(k)}}\underline{1}_{b_l(k)}^T\right)\,;\quad k=1,\ldots,u_l;\ l=1,\ldots,L\,, \end{cases}$$

K_s being a matrix obtained deleting the first row equal to $\frac{1}{\sqrt{s}}\underline{1}_s^T$ of a $s\times s$ orthogonal matrix. Thus, with $\Theta(\underline{h})^+$ and $\Theta(\underline{h})^-$ being the subsets of $\Theta(\underline{h})$ for which $m(\underline{h},\underline{k})$ is respectively even and odd, the generalized F test statistic will be

$$F(\underline{h})=\frac{\displaystyle\sum_{\underline{k}\in\Theta(\underline{h})^+}\frac{s(\underline{k})}{g(\underline{k})}}{\displaystyle\sum_{\underline{k}\in\Theta(\underline{h})^-}\frac{s(\underline{k})}{g(\underline{k})}}\,;\qquad \underline{h}\neq\underline{0}$$

Now, see Fonseca *et al.* (2003a), $s(\underline{k})$ is the product of a central chi-square with $g(\underline{k})$ degrees of freedom by

$$\gamma(\underline{k})=\sigma^2+\sum_{\underline{h}:\underline{k}\leq\underline{h}}b(\underline{h})\sigma^2(\underline{h})\,;\quad \underline{k}\in\Gamma\,.$$

Thus, $F(\underline{h})$ will be the ratio of two linear combinations of central chi-squares with coefficients

$$a(\underline{k}) = \frac{\gamma(\underline{k})}{g(\underline{k})}\,; \quad \underline{k} \in \Gamma\,.$$

Once given this general presentation it may be interesting to consider a simple example. If one factor crosses with a second factor that nests a third factor there will not be, see Khuri *et al.* (1998), pg. 39, an unbiased estimator of the variance component for the second factor given by the difference of two mean squares. To lighten the writting, we replace $\underline{k}$ by (k_1, k_2). Then the first, second and third factors will correspond to the pairs $(1,0)$, $(0,1)$ and $(0,2)$. Likewise, the interactions between the first and the second factors will correspond to the pair $(1,1)$ and the interaction between the first and third factors to $(1,2)$. Now, it is easily seen that

$$\begin{cases} \Theta^{+}(0,1) = \{(0,1),(1,2)\} \\ \Theta^{-}(0,1) = \{(1,1),(0,2)\} \end{cases}$$

and

$$\begin{cases} \gamma(0,1) = \sigma^2 + b(0,1)\sigma^2(0,1) + b(1,1)\sigma^2(1,1) + b(0,2)\sigma^2(0,2) + b(1,2)\sigma^2(1,2) \\ \gamma(1,1) = \sigma^2 + b(1,1)\sigma^2(1,1) + b(1,2)\sigma^2(1,2) \\ \gamma(0,2) = \sigma^2 + b(0,2)\sigma^2(0,2) + b(1,2)\sigma^2(1,2) \\ \gamma(1,2) = \sigma^2 + b(1,2)\sigma^2(1,2) \end{cases}$$

where, with a_1, a_2 and a_3, the number of levels for the three factors,

$$\begin{cases} b(0,1) = a_1 a_3 r \\ b(1,1) = a_3 r \\ b(0,2) = a_1 r \\ b(1,2) = r \end{cases} \quad \text{and} \quad \begin{cases} g(0,1) = a_2 - 1 \\ g(1,1) = (a_1 - 1)(a_2 - 1) \\ g(0,2) = a_2(a_3 - 1) \\ g(1,2) = (a_1 - 1)a_2(a_3 - 1)\,. \end{cases}$$

5. NUMERICAL STUDIES

In this section we study the behavior of simple truncations of the exact distribution, asymptotic distributions with rescaled weights and near-exact distributions described in Section 3, for 4 different simple situations. In order to make things more easily comparable, in all situations we deal with statistics that are the ratio of two linear combinations of two independent Gamma random variables. The two first cases are directly taken from the testing procedure considered in the previous section.

First we consider the case of a generalized F statistic that was used by Fonseca *et all* (2003b) in a study of grapevine clones. In this experiment three clones from two origins were compared. The factors considered were: location of the experiment, origin, and clone. The first factor crosses with the second that nests the third. Since there is no unbiased estimator for the variance component of the second factor (Fonseca *et all.*, 2003a), a generalized F test was used for the corresponding nulity hypothesis, as described in the previous section.

In this case the coefficients of the linear combination of two independent chi-squared random variables in the numerator are much different, with a ratio around 180, while the raio of the two coefficients of the linear combination of two chi-squared random variables in the denominator is only around 2. Although in our numerical studies we have chosen to use always $m_1 = m_2$, in this case we would only have to use quite large values for m_1.

In fact, it happens that the distribution of the sum of independent Gamma random variables with different rate parameters, on which the distributions of statistics that are the ratio of two independent linear combinations of Gamma random variables is based, are mixtures, whose weights decrease much slower when there is unbalance in the values of the ratios of the rate parameters by the coefficients of the Gamma distributions in the linear combination.

That the values of m_1 and m_2 we have to consider in order to obtain a reasonable approximation are really function of the ratios $(\lambda_1/a_1)/(\lambda_2/a_2)$ and $(\lambda_1^*/b_1)/(\lambda_2^*/b_2)$, being larger when such ratios get away from 1, may be stressed since in this first case we would have obtained exactly the same results if we would have used for m_1 the values in Table 2 and $m_2 = 25$.

In fact, in any case, the value of m_1 necessary to attain a reasonable approximation will be a function of the magnitude of the ratios

$$\max_{1\leq i\leq n_1} \mu_1/\mu_i \qquad \text{and} \qquad \min_{1\leq i\leq n_1} \mu_1/\mu_i$$

where, for $i = 1, \ldots, n_1$,

$$\mu_i = \lambda_i/a_i \qquad \text{and} \qquad \mu_1 = \max_{1\leq i\leq n_1} \mu_i,$$

with larger values of m_1 associated with values of the above ratios far from 1, while the values of m_2 are a function of the ratios

$$\max_{1\leq i\leq n_2} \mu_1^*/\mu_i^* \qquad \text{and} \qquad \min_{1\leq i\leq n_2} \mu_1^*/\mu_i^*$$

where, for $i = 1, \ldots, n_2$,

$$\mu_i^* = \lambda_i^*/b_i \qquad \text{and} \qquad \mu_1^* = \max_{1\leq i\leq n_2} \mu_i^*,$$

once again with larger values of m_2 associated with values of the above ratios far from 1.

The second studied case is derived from the first one just by using coefficients for the linear combination of chi-squared random variables in the numerator and in the denominator that yield exactly ratio of 3/2 and 2, respectively.

In the third and fourth cases we use ratios of two linear combinations of Gamma random variables to show that our approach also applies to this case.

In Table 1 below are summarized the values of the parameters used in the 4 cases studied.

Table 1. – Parameters for the 4 cases studied (all with $n_1 = n_2 = 2$)

	a_1	a_2	s_1	s_2	λ_1	λ_2	b_1	b_2	s_1^*	s_2^*	λ_1^*	λ_2^*
Case 1	a_1^*	a_2^*	1/2	4	1/2	1/2	b_1^*	b_2^*	1	2	1/2	1/2
Case 2	3/2	1	1/2	4	1/2	1/2	1	2	1	2	1/2	1/2
Case 3	3/2	1	7/3	23/2	9/2	9/2	1	2	9/4	5/2	19/2	19/2
Case 4	7/2	1	7/3	23/2	9/2	9/2	1	2	9/4	5/2	19/2	19/2

a_1^*=9.61163×10^{0} a_2^*=5.33912×10^{-2}
b_1^*=8.61565×10^{-1} b_2^*=3.91407×10^{-1}

In Tables 2 through 5 we may see the values of the tail probabilities relative to the exact 0.95 and 0.99 quantiles for each of the cases considered, for different values of $m_1 = m_2$ used in the truncations mentioned in section 3.

We may see how in each case the simple truncation is not an adequate approximation to the exact distribution as soon as the number of terms used starts to decline and namely when we try to use a rather small number of terms. The asymptotic approximations based on truncations with rescaled weights behave much better than the simple truncations. Anyway, in every case the near-exact distributions have an even much better performance.

Although for this first case studied the exact distribution is known to have a concise finite representation (Fonseca *et all.*, 2003b) we used it to illustrate that in some cases we may need a fairly large number of terms in the summations in order to obtain an accetable quality of the approximation. This is due to the fact that in this case the ratio of the values of the coefficients used in the linear combination in the numerator is far from 1.

Table 2. – Case 1 tail probabilities for the exact truncated (trunc), asymptotic with rescaled weights (asymp) and near-exact (near-ex) distributions

	quantile 0.95 values of $m_1 = m_2$					
	500	300	250	200	150	120
trunc	0.943791	0.911639	0.889635	0.854890	0.799441	0.750602
asymp	0.961267	0.977296	0.982637	0.988069	0.993090	0.995641
near-ex	0.952083	0.950755	0.948678	0.945255	0.940287	0.936649
	quantile 0.99 values of $m_1 = m_2$					
	500	300	250	200	150	120
trunc	0.975695	0.930012	0.903401	0.864017	0.804407	0.753552
asymp	0.993761	0.996992	0.997842	0.998617	0.999258	0.999554
near-ex	0.988863	0.985962	0.984932	0.983935	0.983105	0.982884

In this first case we may also see how the asymptotic approximation with rescaled weights although having a quite good performance for the 0.99 quantile, does have a quite poor performance for the 0.95 quantile, what stresses the importance of the near-exact approximation, which has the best performance for both quantiles and for any of the values of m_1 and m_2 used.

Table 3. – Case 2 tail probabilities for the exact truncated (trunc), asymptotic with rescaled weights (asymp) and near-exact (near-ex) distributions

	quantile 0.95 values of $m_1 = m_2$					
	25	20	15	10	5	3
trunc	0.950000	0.949995	0.949863	0.946826	0.887249	0.760113
asymp	0.950000	0.950000	0.949993	0.949841	0.946746	0.939234
near-ex	0.950000	0.950000	0.950000	0.950000	0.950079	0.950170
	quantile 0.99 values of $m_1 = m_2$					
	25	20	15	10	5	3
trunc	0.990000	0.989995	0.989863	0.986825	0.927180	0.799456
asymp	0.990000	0.990000	0.989999	0.989968	0.989354	0.987847
near-ex	0.990000	0.990000	0.990000	0.990000	0.990023	0.990162

In this second case the near-exact approximation shows a very good performance for both quantiles, even for very small values of m_1 and m_2 for which the asymptotic approximation with rescaled weights starts to perform not so well.

Table 4. – Case 3 tail probabilities for the exact truncated (trunc), asymptotic with rescaled weights (asymp) and near-exact (near-ex) distributions

	quantile 0.95 values of $m_1 = m_2$					
	25	20	15	10	5	3
trunc	0.949999	0.949986	0.949683	0.943649	0.842673	0.647008
asymp	0.950000	0.949999	0.949984	0.949696	0.945279	0.936277
near-ex	0.950000	0.950000	0.950000	0.950016	0.950595	0.950749
	quantile 0.99 values of $m_1 = m_2$					
	25	20	15	10	5	3
trunc	0.989999	0.989986	0.989683	0.983637	0.881733	0.682278
asymp	0.990000	0.990000	0.989997	0.989940	0.989095	0.987316
near-ex	0.990000	0.990000	0.990000	0.990004	0.990266	0.991032

In the fourth case we will have to use a larger number of terms to take care of the greater unbalance of the ratios of the coefficients in the linear combination of the numerator.

Table 5. – Case 4 tail probabilities for the exact truncated (trunc), asymptotic with rescaled weights (asymp) and near-exact (near-ex) distributions

	quantile 0.95 values of $m_1 = m_2$					
	50	30	20	10	5	3
trunc	0.950000	0.949698	0.942750	0.821178	0.489308	0.257356
asymp	0.950001	0.950188	0.952014	0.962322	0.970567	0.970832
near-ex	0.950000	0.949979	0.949458	0.944542	0.940024	0.939144
	quantile 0.99 values of $m_1 = m_2$					
	50	30	20	10	5	3
trunc	0.989999	0.989543	0.980956	0.847435	0.501586	0.263787
asymp	0.990000	0.990066	0.990596	0.993092	0.994922	0.995095
near-ex	0.990000	0.989977	0.989803	0.989795	0.991384	0.992668

In both these two last cases the simple truncation of the exact distribution shows a very poor performance, mainly for lower number of terms while the asymptotic approximation with rescaled weights displays an acceptable behavior as long as the number of terms remains moderately large. Once again, the near-exact approximation shows an outstanding performance even for quite small number of terms.

6. CONCLUDING REMARKS

We have shown how through the use of near-exact approximations that equate the two first moments of the exact distribution we were able to obtain approximations that almost coincide with the exact distribution. This fact, largely overcomes the minor drawback that is the requirement of the existence of the two first moments of the generalized F statistics. Indeed, according to (10), this happens whenever $r_2 = \sum_{i=0}^{n_2} s_i^* > 2$, which is not a very limitative requirement.

REFERENCES

Fonseca, M., Mexia, J. T. & Zmyslony, R. (2002). Exact distribution for the generalized F tests, *Discussiones Mathematicae, Probability and Statistics*, **22**, 37-51.

Fonseca, M., Mexia, J. T. & Zmyslony, R. (2003a). Estimators and tests for variance components in cross nested orthogonal models. *Discussiones Mathematicae, Probability and Statistics*, **23**, 173-201.

Fonseca, M., Mexia, J. T. & Zmyslony, R. (2003b). Estimating and testing of variance components: an application to a grapevine experiment. *Biometrical Letters*, **40**, 1-7.

Khuri, A. I., Mathew, T. & Sinha, B. K. (1998). *Statistical Tests for Mixed Linear Models*, J. Wiley, New York.

Michalsky, A. & Zmyślony, R. (1996). Testing hypotheses for variance components in mixed linear models, *Statistics*, **27**, 297-310.

Michalsky, A. & Zmyślony, R. (1999). Testing hypotheses for linear functions of parameters in mixed linear models, *Tatra Mountain Mathematical Publications*, **17**, 103-110.

Moschopoulos, P. G. (1985). The distribution of the sum of independent Gamma random variables, *Ann. Inst. Statisit. Math.*, **37**, 541-544.

Nunes, C. & Mexia, J. T. (2004). Non-central generalized F distributions, *Discussiones Mathematicae, Probability and Statistics* (accepted for publication).