СТОПАНСКА АКАДЕМИЯ

„Димитър А. Ценов“ - Свищов

Международен научен форум

Световни и национални бизнес трансформации – проактивни управленски, финансово-счетоводни и планови решения

25-26 октомври 2024 г.

Сборник с доклади

Стопанска академия „Д. А. Ценов“
гр. Свищов, Република България

D. A. TSENOV ACADEMY OF ECONOMICS – SVISHTOV

INTERNATIONAL SCIENTIFIC FORUM

GLOBAL AND NATIONAL BUSINESS TRANSFORMATIONS – PROACTIVE MANAGEMENT, FINANCIAL-ACCOUNTING AND PLANNING SOLUTIONS

Svishtov, 25-26 October 2024

Conference Proceedings

Tsenov Academic Publishing House
Svishtov
2024

ISBN (print) **978-954-23-2513-0**
ISBN (online) **978-954-23-2514-7**

NEAR-EXACT DISTRIBUTIONS: UNDERSTANDING THEIR IMPORTANCE AND APPLICATIONS

Full Professor Carlos A. Coelho, PhD[1]

Abstract: *Most Likelihood Ratio Tests in Multivariate Analysis have statistics with exact distributions too complicated to be used in their exact form. This is why asymptotic distributions were developed. However, usually, these distributions do not exhibit the required precision. While the commonly used chi-square approximations fall too far from the exact distribution, even for large sample sizes, other asymptotic distributions commonly used not only worsen their performance for increasing numbers of variables but also are no longer proper distributions when the ratio between the sample size and the number of variables approaches 1. This inadequate behavior is not easy to overcome, when common asymptotic techniques are used to develop these distributions. But, using a different approach, factorizing the characteristic function of the statistic under study, and replacing then only a smaller part of it by an adequate asymptotic approximation, it is possible to build what we call 'near-exact' distributions. These fall extremely close to the exact distribution, even for very small samples, and exhibit an asymptotic behavior also for increasing number of variables. In areas as Business Planning and Forecasting, where usually we need to deal with large numbers of variables, the use of near-exact distributions may thus be of great importance in order to be able to take the right decisions, based on accurate testing procedures.*

Keywords: *characteristic function, near-exact distributions, likelihood ratio statistics, Multivariate Analysis, quantiles, p-values*

JEL: C44, C46

1. Introduction

Likelihood Ratio Test (LRT) statistics used in Multivariate Analysis have usually quite complicated exact distributions, whose p.d.f.'s (probability density functions) and c.d.f.'s (cumulative distribution functions) do not have manageable expressions, thus requiring the use of approximations.

The most used and widespread asymptotic approximations for the distributions of these statistics, or more precisely, for the distribution of the negative logarithm of these statistics, are the chi-square asymptotic approximation (Wilks, 1938; Anderson, 2003, Sec. 8.8, 12.4.1; Muirhead, 2005, Thms. 10.7.5, 11.3.9, Sec. 11.2.4), and asymptotic distributions based on the results in the paper by Box (1949). These are seen by many authors as a very useful tool (Gleser & Olkin, 1975; Brunner et al., 1977; Anderson, 2003, Secs. 8.5.1, 8.5.2, 9.4, 10.5; Muirhead, 2005, Sec. 8.2.4, 11.2.4).

It is the aim of this note to bring to the attention of the reader that although the single chi-square asymptotic approximation to the distribution of $-2\,log\,\Lambda$, (where Λ represents the LRT statistic), based on increasing sample sizes, is indeed

[1] Mathematics Department, NOVA School of Science and Technology, and NOVA Math – Center for Mathematics and Applications, NOVA University of Lisbon; email: cmac@fct.unl.pt

a legitimate one, derived from convergence in distribution results, it does not yield the desired precision not even for quite large sample sizes, giving too low quantiles and percentage points, leading thus to spurious rejections of the null hypothesis, a problem that has been largely overlooked. However, that this single chi-square asymptotic approximation may not yield the desired results has already been recognized by some authors (Brunner et al., 1977). It is also the aim of this note to bring to the reader's attention that in situations where the ratio between the sample size and the number of variables tends to 1, Box asymptotic distributions usually not yield legitimate distributions, with the „p.d.f." showing values below zero and the „c.d.f." with values below zero and sometimes also above 1, a fact that seems to have never been noticed by other authors. Of course, this fact entails that, mainly in situations where the sample sizes are not that large, the quantiles and p-values given by these distributions fall way off from the exact ones.

Furthermore, it is more or less a known fact that these asymptotic approximations worsen their performance when the number of variables increases, which is a rather unwelcome feature, moreover since nowadays with the great ease in collecting and storing data, the number of variables used is most often rather large. One other inconvenient feature of these asymptotic distributions is that they perform quite bad for small sample sizes. But, one even worse and more unwelcome feature of these asymptotic distributions, which has been completely overlooked by other authors, is that these asymptotic „distributions" are no longer proper distributions in situations where the sample sizes exceed the number of variables only moderately, thus giving in these cases erroneous *p*-values and quantiles. A problem that also gets worse for increasing numbers of variables.

But these problems are not easy to overcome when we use the common asymptotic techniques to develop asymptotic distributions.

2. Near-Exact distributions

It was to overcome the above mentioned problems that occur with the common asymptotic distributions that the so-called 'near-exact distributions' were developed. These are asymptotic distributions obtained by first decomposing the characteristic function (c.f.) of the statistic under study, or of its negative logarithm, decomposition that in case of the LRT statistics used in Multivariate Analysis is more often a factorization of the type

$$\Phi_W(t) = \Phi_{W,1}(t)\,\Phi_{W,2}(t)\,,$$

where $W = -\,log\,\Lambda$ is the negative logarithm of the LRT statistic Λ, and $\Phi_W(t)$ is its c.f., and where $\Phi_{W,1}(t)$ represents the part of $\Phi_W(t)$ that corresponds to a known manageable distribution and $\Phi_{W,2}(t)$ represents the part of $\Phi_W(t)$ that corresponds to a non-manageable distribution. In most cases it is even possible, by a clever handling of $\Phi_W(t)$, to define $\Phi_{W,1}(t)$ and $\Phi_{W,2}(t)$ in such a way that

$\Phi_{W,1}(t)$ corresponds to 'the major part' of $\Phi_W(t)$, that is, in such a way that

$$\int_{-\infty}^{+\infty} \left|\Phi_W(t) - \Phi_{W,1}(t)\right| dt \ll \int_{-\infty}^{+\infty} \left|\Phi_W(t) - \Phi_{W,2}(t)\right| dt \,.$$

Then we will keep $\Phi_{W,1}(t)$ untouched and replace $\Phi_{W,2}(t)$ by an asymptotic adequate result, say $\Phi^*_{W,2}(t)$, in such a way that $\Phi_{W,1}(t)\Phi^*_{W,2}(t)$ corresponds to a known manageable distribution, from which quantiles and p-values may be easily obtained.

This technique has been successfully used to obtain near-exact distributions for many LRT statistics used in Multivariate Analysis (Coelho et al., 2010; Marques et al., 2011; Coelho & Marques, 2013; Coelho, 2021; Coelho and Pielaszkiewicz, 2021).

Most times these near-exact distributions assume the form of finite mixtures of GNIG (Generalized Near-Integer Gamma) distributions (Coelho, 2004), with p.d.f.'s and c.d.f.'s for $W = -log\,\Lambda$ of the form

$$\sum_{k=0}^{m^*} \pi_k \; f^{GNIG}\left(w \,\middle|\, \{r_j\}_{j=1,\ldots,p^*}, r; \{\lambda_j\}_{j=1,\ldots,p^*}, \lambda; p^*+1\right) \qquad (1)$$

and

$$\sum_{k=0}^{m^*} \pi_k \; F^{GNIG}\left(w \,\middle|\, \{r_j\}_{j=1,\ldots,p^*}, r; \{\lambda_j\}_{j=1,\ldots,p^*}, \lambda; p^*+1\right) \qquad (2)$$

where $f^{GNIG}(\cdot)$ and $F^{GNIG}(\cdot)$ represent respectively the p.d.f. and the c.d.f. of the GNIG distribution (see for example Appendix C in Coelho (2021) for the notation used) and where the weights π_k $(k = 0, \ldots, m^*)$ are computed in such a way that the first m^* exact moments of W are matched by the near-exact distribution, and then with $\pi_{m^*} = 1 - \sum_{k=0}^{m^*-1} \pi_k$, so that $\sum_{k=0}^{m^*} \pi_k = 1$. In (1) and (2) the r_j $(j = 1, \ldots, p^*)$ and r are the shape parameters of the GNIG distribution and the λ_j $(j = 1, \ldots, p^*)$ and λ are the rate parameters. In case r is an integer, then the GNIG distribution converts to a GIG distribution (Coelho, 1998).

3. The LRT for equality of covariance matrices

As an example of an LRT for which it is important to build near-exact distributions, since it is an LRT for which statistic it is not possible to obtain its exact distribution, or rather, its p.d.f. or c.d.f., in a finite closed form, we use the LRT for equality of covariance matrices. Near-exact distributions for this LRT statistic were developed in Coelho et al. (2010), Marques et al. (2011) and Coelho & Marques (2012).

Although we do not have the expression for the exact c.d.f. of the LRT statistic, neither for its logarithm, we can obtain a very sharp upper bound on the absolute value of the difference between this c.d.f. and that of any approximate distribution using the measure

$$\Delta = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\left|\frac{\Phi_W(t)-\Phi_W^*(t)}{t}\right|dt$$

with

$$\Delta \geq \max_{w>0}|F_W(w)-F_W^*(w)| = \max_{0<z<1}|F_\Lambda(z)-F_\Lambda^*(z)|\,,$$

where $\Phi_W(t)$ is the exact c.f. of W, $\Phi_W^*(t)$ its approximate (near-exact or asymptotic) c.f., and $F_W(t)$ and $F_\Lambda(z)$ the exact c.d.f.'s of W and Λ, and $F_W^*(w)$ and $F_\Lambda^*(z)$ their approximate c.d.f.'s.

Table 1 – Values of the measure Δ for the approximating distributions for the LRT statistic to test equality of q covariance matrices

			near-exact number of exact moments matched			Box-And	Chi-square
p	*q*	*n*	4	6	10		
5	6	7	2.20×10^{-8}	1.02×10^{-10}	5.41×10^{-15}	6.15×10^{-1}	1.08×10^{0}
		55	2.89×10^{-13}	2.10×10^{-17}	5.74×10^{-25}	4.69×10^{-2}	1.42×10^{-1}
	12	7	1.20×10^{-8}	4.41×10^{-11}	1.31×10^{-15}	8.21×10^{-1}	1.22×10^{0}
		55	2.51×10^{-13}	1.27×10^{-17}	7.87×10^{-26}	6.91×10^{-2}	1.98×10^{-1}
20	6	22	1.70×10^{-13}	6.57×10^{-18}	1.81×10^{-26}	1.46×10^{0}	1.64×10^{0}
		70	2.95×10^{-15}	3.08×10^{-20}	6.04×10^{-30}	1.52×10^{-1}	9.19×10^{-1}
	12	22	5.96×10^{-15}	5.28×10^{-20}	5.77×10^{-30}	1.95×10^{0}	1.77×10^{0}
		70	1.42×10^{-16}	4.03×10^{-22}	4.30×10^{-33}	2.24×10^{-1}	1.06×10^{0}
50	6	52	7.24×10^{-15}	6.35×10^{-20}	6.68×10^{-30}	3.16×10^{0}	2.01×10^{0}
		100	5.83×10^{-15}	4.92×10^{-20}	4.63×10^{-30}	4.74×10^{-1}	1.54×10^{0}
		500	8.36×10^{-19}	3.55×10^{-25}	8.46×10^{-38}	4.71×10^{-2}	7.74×10^{-1}
	12	52	1.50×10^{-16}	2.61×10^{-22}	9.90×10^{-34}	4.40×10^{0}	2.14×10^{0}
		100	1.18×10^{-16}	2.14×10^{-22}	8.59×10^{-34}	7.72×10^{-1}	1.67×10^{0}
		500	1.58×10^{-20}	1.43×10^{-27}	1.39×10^{-41}	6.97×10^{-2}	9.24×10^{-1}

In Table 1 are the values of Δ for the near-exact distributions in Coelho et al. (2010), the chi-square asymptotic distribution for $-2\,log\,\Lambda$, which for W takes the form of a Gamma distribution, and the Box asymptotic distribution in section 10.5 of Anderson (2003), designated as Box-And in the table. In this table, as well as in Fig. 1, p, n and q represent respectively the number of variables, the sample size and the number of matrices being tested.

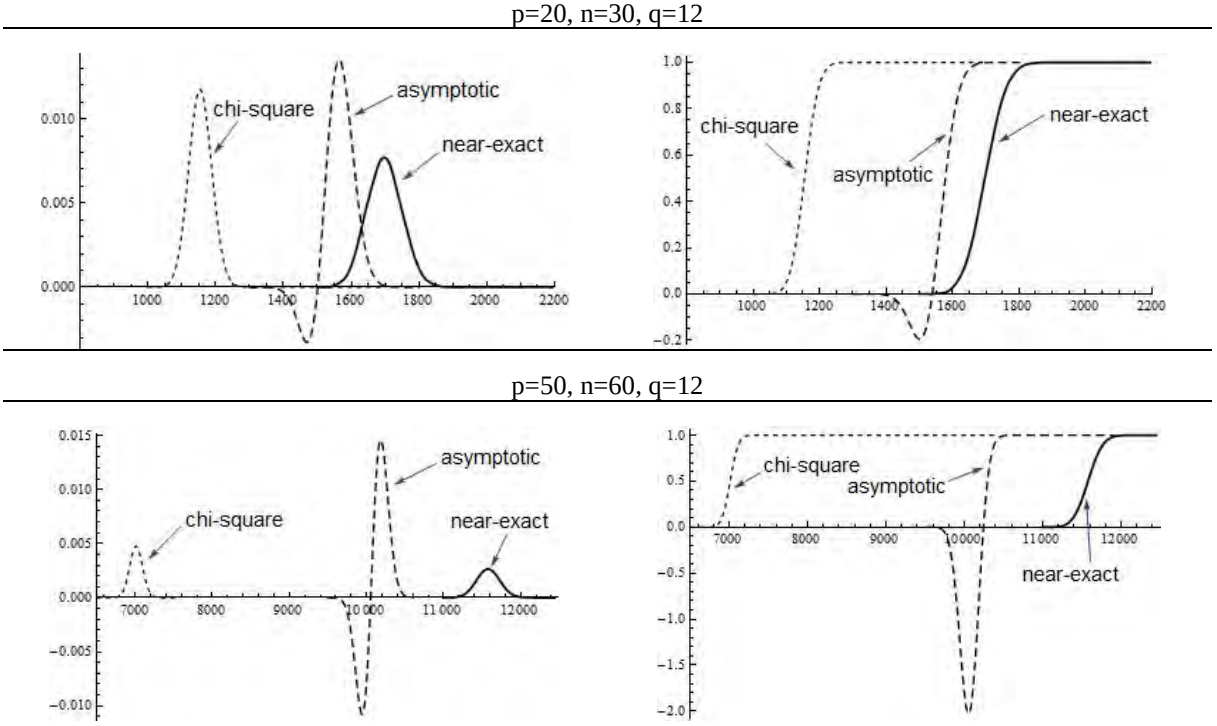

Figure 1 – Plots of asymptotic and near-exact p.d.f.'s (left) and c.d.f.'s (right) of the negative logarithm of the LRT statistic to test equality of several covariance matrices

We can see from Table 1 the extremely low values for the measure Δ for the near-exact distributions, even for the smaller values of n and the smaller values of m^*, that is, for the smaller sample sizes and for the smaller number of exact moments of W matched by the near-exact distributions, which show the extreme closeness of these distributions to the exact distribution. Also, the quite large values of Δ, some of them even larger than 1, for some of the asymptotic distributions, namely for smaller values of n, show that these distributions either are not legitimate distributions, as it happens with some of the Box asymptotic distributions, or they fall way too far from the exact distribution. These facts may also be analyzed by looking at the plots in Figure 1.

References

Anderson, T. W. (2003) *An Introduction to Multivariate Statistical Analysis*, J. Wiley & Sons, Hoboken, New Jersey.

Box, G. E. P. (1949) „A general distribution theory for a class of likelihood criteria", *Biometrika*, 36, 317–346.

Brunner, E., Dette, H. & Munk, A. (1977). Box-Type Approximations in Nonparametric Factorial Designs, *J. Amer. Stat. Assoc.*, 92, 1494-1502.

Coelho, C. A. (1998). The Generalized Integer Gamma Distribution – a Basis for Distributions in Multivariate Statistics. *J. Multiv. Analysis*, 64, 86-102.

Coelho, C.A. (2004) „The Generalized Near-Integer Gamma distribution: a basis for near-exact

approximations to the distributions of statistics which are the product of an odd number of independent Beta random variables“, *J. Multiv. Analysis*, 89, 191–218.

Coelho, C. A. (2021). Testing equality of mean vectors with block-circular and block compound-symmetric covariance matrices, in ‘Multivariate, Multilinear and Mixed Linear Models’, Filipiak, K., Markiewicz, A., von Rosen, D. (eds.), 157–201, Springer series in Contributions to Statistics.

Coelho, C. A. & Marques, F. J. (2012). Near-exact distributions for the likelihood ratio test statistic to test equality of several variance-covariance matrices in elliptically contoured distributions, *Comput. Statist.*, 27, 627–659.

Coelho, C. A. & Marques, F. J. (2013). The Multi-Sample Block-Scalar Sphericity Test: Exact and Near-Exact Distributions for Its Likelihood Ratio Test Statistic, *Comm. Statist. Theory Methods*, 42, 1153–1175.

Coelho, C. A. & Pielaszkiewicz, J. (2021). The Likelihood Ratio Test of Equality of Mean Vectors with a Doubly Exchangeable Covariance Matrix, in ‘Methodology and Applications of Statistics – A Volume in Honor of C. R. Rao on the Occasion of his 100th Birthday’, Arnold, B., Balakrishnan, N., Coelho, C. A. (eds.), Springer series in Contributions to Statistics, 151–191.

Coelho, C. A., Marques, F. J. & Arnold, B. C. (2010). Near-exact distributions for certain likelihood ratio test statistics, *J. Stat. Theory Pract.*, 4, 711–725.

Gleser, L. J. & Olkin, I. (1975). A note on Box’s general method of approximation for the null distributions of likelihood criteria, *Ann. Inst. Statist. Math.*, 27, 319–326.

Marques, F. J., Coelho, C. A. & Arnold, B. C. (2011). A general near-exact distribution theory for the most common likelihood ratio test statistics used in Multivariate Analysis. *TEST*, 20, 180–203.

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory*, 2nd ed., J. Wiley & Sons, Hoboken, New Jersey.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math.Stat.*, 9, 60–62.