Near-exact distributions: what are they and why do we need them?

Carlos A. Coelho

Departamento de Matemática and Centro de Matemática
e Aplicações, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, Portugal
cmac@fct.unl.pt

Abstract

Indeed, in multivariate analysis, most of the commonly used asymptotic distributions worsen their performance when the number of variables increase and even many of them are no longer proper distributions when the number of variables goes above a given threshold. These are facts that have been completely overlooked by other authors and this awkward behavior is not easy to overcome, when we use the common asymptotic techniques. However, by using a different approach, which combines an adequate decomposition of the characteristic function of the statistic under study, most often a factorization, with the action of keeping then most of this characteristic function unchanged, and replacing the remaining smaller part by an adequate asymptotic approximation, it is possible to build manageable approximations, called 'near-exact' approximations, which yield distributions extremely close to the exact distribution, and which exhibit a very good performance for very small sample sizes and an asymptotic behavior not only for increasing sample sizes but also for increasing number of variables involved. These near-exact distributions may then be applied to obtain very well-fitting near-exact quantiles and *p*-values and they have been, so far, successfully applied to a large number of statistics. Examples are given.

Keywords: characteristic function, distribution of likelihood ratio statistics, Fox *H* function.

1 Introduction

In this short paper the author is going to address the issue of approximating the distribution of several l.r.t. (likelihood ratio test) statistics used in Multivariate Analysis. These statistics 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 these statistics, are the asymptotic distributions based on Box (1949) paper and they are seen by many authors as a very useful general tool (Gleser and Olkin, 1975; Anderson, 2003, Chaps. 8, 9, 10). However, it is more or less a known fact that these approximations worsen their performance when the number of variables increases, which is a rather embarrassing feature, moreover since nowadays with the great ease in collecting and storing data, the number of variables used may be rather large. One other inconvenient feature of these asymptotic distributions is that they perform quite bad for very small sample sizes, that is, sample sizes that barely exceed the number of variables. But, one even worse and more embarrassing feature of these asymptotic distributions, which have been completely overlooked by all authors, is that these asymptotic "distributions" are no longer proper distributions for moderately large numbers of variables involved and small to moderate sample sizes, thus giving in these cases erroneous *p*-values and quantiles.

These facts may be checked by using a measure of distance between distributions based on the c.f.'s (characteristic functions). Since usually for any l.r.t. statistic Λ we

are able to obtain the expression for its h-th exact moment and usually these expressions remain valid for any complex h, if we take $W = -\log \Lambda$, we will be able to easily obtain its exact c.f. as

$$\Phi_W(t) = E\left(e^{\mathrm{i}tW}\right) = E\left(e^{-\mathrm{i}t\log\Lambda}\right) = E\left(\Lambda^{-\mathrm{i}t}\right),$$

although most of times the exact p.d.f. and c.d.f. are not obtainable in a closed manageable form.

Then we may use the measure

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

where $\Phi_W(t)$ is the exact c.f. of W and $\Phi_W^*(t)$ represents the c.f. corresponding to the approximate distribution of W under study, as a measure of proximity between the exact and that approximate distribution of W.

This measure is related with the Berry-Esseen upper bound (Hwang, 1998; Loève, 1977) and was already used in several papers to assess the "distance", or rather, the "proximity" between the exact and approximate distributions (Coelho and Marques, 2009, 2010, 2012, 2013; Coelho et al., 2010) and it provides a sharp and useful upper bound on the difference between the exact and approximate c.d.f.'s, since

$$\Delta \geq \max_{w\in S_W}\left|F_W(w) - F_W^*(w)\right| = \max_{\ell\in S_\Lambda}\left|F_\Lambda(\ell) - F_\Lambda^*(\ell)\right|,$$

where S_W and S_Λ represent the supports of W and Λ, $F_W(\cdot)$ and $F_W^*(\cdot)$, respectively the exact c.d.f. of W and the c.d.f. corresponding to $\Phi_W^*(t)$, while $F_\Lambda(\cdot)$ and $F_\Lambda^*(\cdot)$ are the exact and approximate c.d.f.'s of Λ.

For proper distributions we have $0 < \Delta < 1$, while for not proper distributions, in most cases we have $\Delta > 1$, since for these "distributions" the "c.d.f." usually goes both below zero as well as above 1. Numerical studies using the measure Δ, may be analyzed in Section 3.

2. The proposed solution for the problem — the near-exact distributions

Given the rather complex structure of the exact distributions of most of the l.r.t. statistics used in Multivariate Analysis, the solution proposed consists in identifying: i) a part of the exact distribution that we are able to handle and which should be left unchanged, and, ii) the remaining part, which has to be asymptotically approximated. This has to be done in such a way that the resulting final distribution is manageable, in the sense that we will be able to obtain a manageable c.d.f. from which it will be easy to compute p-values and quantiles.

This is usually achieved by working on the c.f. of W, the negative logarithm of the l.r.t. statistic, which is usually easy to obtain, as we remarked in the previous section. This work consists usually in obtaining an adequate factorization of the c.f. of W, identifying then the terms which should be left unchanged and those which we should asymptotically approximate. That is, if we are able to write $\Phi_W(t)$ as

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

where $\Phi_{1,W}(t)$ encompasses all the terms to be left unchanged and $\Phi_{2,W}(t)$ all the terms to be asymptotically approximated, then we will approximate $\Phi_{2,W}(t)$ by $\Phi_{2,W}^*(t)$, so that

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

will become what we call a near-exact c.f. of W. Of course, in order for the whole process to be useful, $\Phi_W^*(t)$ has to correspond to a manageable distribution, from which p-values and quantiles can be easily computed.

Nevertheless how much complicated this whole process may seem to be, in practice, the process of building near-exact distributions is not so complicated, and so far near-exact distributions have already been built for a large array of l.r.t. statistics used in Multivariate Analysis (Coelho, 2004; Coelho and Marques, 2010, 2012; Coelho et al., 2010) and they may be even easily developed for tests of quite complicated structures of the covariance matrices being tested, by considering the decomposition of the null hypothesis into a set of conditionally independent hypotheses (Coelho and Marques, 2009, 2013).

Furthermore, these near-exact distributions also have the much welcome features of, besides being asymptotic for increasing sample sizes, being also asymptotic for increasing numbers of variables involved and also, in the multi-sample cases, for increasing number of populations involved, besides having amazing performances for very small sample sizes, which even improve as the number of variables increases.

3. Some examples and numerical studies

In this section we consider the l.r.t. (likelihood ratio test) statistics Λ used to test: i) independence of sets of variables, ii) sphericity of the covariance matrix, and iii) equality of q covariance matrices. Throughout, p denotes the number of variables and n the sample size.

The near-exact distributions used are the ones developed in Coelho et al. (2010), matching 4, 6 and 10 exact moments, with a slight change in the computation of the parameter r, which leads to even better performing approximations. The asymptotic distributions considered are the common chi-square approximation to the distribution of $-2 \log \Lambda$ (Anderson, 2003, Chaps. 8,9,10), denoted by 'Chi-square' and the Box-style asymptotic distributions in Chapters 8, 9 and 10 of Anderson (2003), denoted by 'Box-And'. All distributions are reported to $W = -\log \Lambda$.

In Tables 3.1, 3.3 and 3.5 we have the values of the measure Δ in (1) for the near-exact distributions (matching 4, 6 and 10 exact moments), the Box-Anderson asymptotic distribution and the chi-square approximation, for each one of the three l.r.t. statistics named above. From these tables we may see how, opposite to the asymptotic distributions, the near-exact distributions exhibit a clear asymptotic behavior for increasing numbers of variables involved, besides a very good performance for very small sample sizes, with values of the measure Δ in (1) which are millions of times smaller than the ones obtained for the asymptotic distributions.

In Figures 3.1-3.3 we have plots of p.d.f.'s and c.d.f.'s of the near-exact and asymptotic distributions for selected values of p and n (and eventually q). The near-exact distributions used are the ones that match 4 exact moments. We may see how the Box-Anderson asymptotic distributions are no longer proper distributions for just moderately large numbers of variables and small to moderately large sample sizes.

Table 3.1 – Values of the measure Δ for the approximating distributions for the l.r.t. statistic to test independence of sets of variables

		near-exact				
		number of exact moments matched			Box-And	Chi-square
p_k	n	4	6	10		
{3, 3, 4}	15	1.84×10^{-11}	6.67×10^{-14}	1.04×10^{-18}	1.68×10^{-2}	7.76×10^{-1}
	60	8.91×10^{-15}	3.21×10^{-18}	3.02×10^{-24}	3.20×10^{-5}	1.73×10^{-1}
{10, 13, 7}	35	5.80×10^{-17}	2.56×10^{-21}	4.17×10^{-29}	3.66×10^{-1}	1.36×10^{0}
	80	4.09×10^{-18}	8.80×10^{-23}	3.75×10^{-31}	3.12×10^{-3}	8.19×10^{-1}
{7, 7, 10, 6}	35	5.07×10^{-17}	2.11×10^{-21}	3.04×10^{-29}	3.97×10^{-1}	1.36×10^{0}
	80	3.11×10^{-18}	6.01×10^{-23}	2.06×10^{-31}	4.02×10^{-3}	8.21×10^{-1}
{15, 11, 15, 9}	55	2.53×10^{-19}	1.15×10^{-24}	2.26×10^{-34}	1.01×10^{0}	1.59×10^{0}
	100	8.35×10^{-20}	3.10×10^{-25}	4.09×10^{-35}	3.51×10^{-2}	1.18×10^{0}

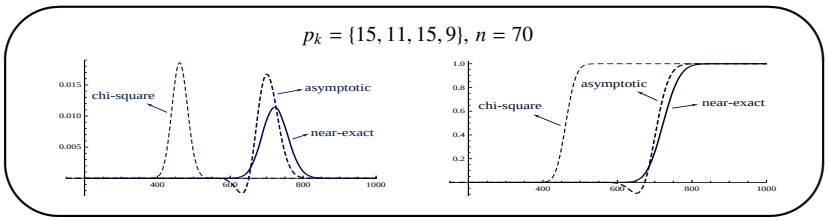

Figure 3.1 – Plots of p.d.f.'s and c.d.f.'s for approximating distributions of the l.r.t. statistic to test independence of sets of variables

Table 3.2 – Quantiles α for the approximating distributions for the l.r.t. statistic to test independence of sets of variables

p_k	n	α	Near-exact	Asymp.	Chi-square
$\{7, 7, 10, 6\}$	40	0.90	374.7008	294.2025	183.2358
		0.95	384.1695	301.5610	188.2775
		0.99	402.4316	315.9025	197.9799
$\{15, 11, 15, 9\}$	70	0.90	768.5490	742.3751	489.7511
		0.95	781.8981	753.2441	497.9141
		0.99	807.3946	774.4247	513.4685

In Tables 3.2, 3.4 and 3.6 we may see the values of some quantiles for the approximating distributions considered. Once again, the near-exact distributions considered were the ones that match 4 exact moments. We may see how the chi-square approximation for $-2 \log \Lambda$, although being valid in terms of convergence in distribution, is indeed of no practical usefulness, given that it lies very much far apart the exact distribution, even for quite large sample sizes, always giving quantiles which are much lower than the exact or near-exact ones, and thus leading to too many spurious rejections of the null hypothesis.

Table 3.3 – Values of the measure Δ for the approximating distributions for the l.r.t. statistic to test sphericity of the covariance matrix

		near-exact number of exact moments matched			Box-And	Chi-square
p	n	4	6	10		
5	10	5.06×10^{-10}	1.30×10^{-11}	1.86×10^{-15}	3.84×10^{-2}	3.92×10^{-1}
	55	1.27×10^{-13}	3.97×10^{-17}	2.37×10^{-23}	7.50×10^{-4}	5.97×10^{-2}
10	15	1.63×10^{-14}	4.93×10^{-17}	2.51×10^{-22}	9.40×10^{-2}	7.48×10^{-1}
	60	2.91×10^{-16}	6.69×10^{-21}	1.58×10^{-29}	2.46×10^{-3}	1.67×10^{-1}
30	35	8.40×10^{-18}	2.40×10^{-23}	4.38×10^{-33}	5.93×10^{-1}	1.32×10^{0}
	80	1.31×10^{-18}	2.11×10^{-24}	5.36×10^{-35}	1.96×10^{-2}	7.89×10^{-1}
50	55	9.58×10^{-20}	3.97×10^{-26}	4.24×10^{-38}	1.22×10^{0}	1.56×10^{0}
	100	5.18×10^{-20}	1.73×10^{-26}	2.09×10^{-38}	8.21×10^{-2}	1.15×10^{0}

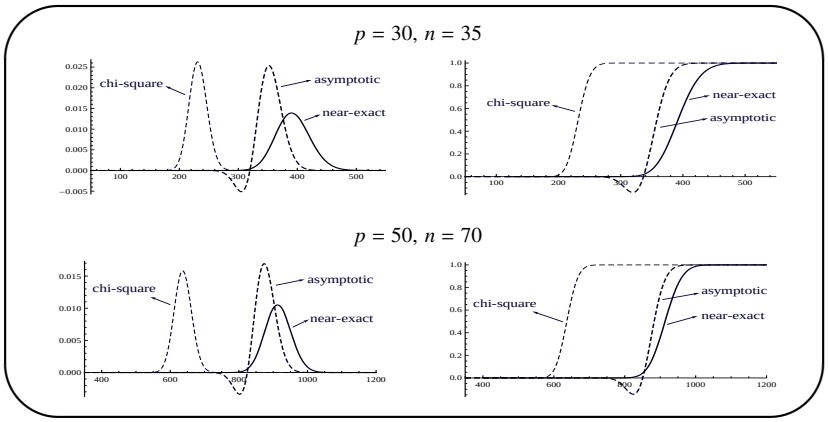

Figure 3.2 – Plots of p.d.f.'s and c.d.f.'s for approximating distributions of the l.r.t. statistic to test sphericity of the covariance matrix

Table 3.4 – Quantiles α for the approximating distributions for the l.r.t. statistic to test sphericity of the covariance matrix

p	n	α	Near-exact	Asymp.	Chi-square
30	35	0.90	430.4899	382.2868	251.7216
		0.95	441.9999	389.9534	257.6092
		0.99	464.3932	405.0379	268.8972
50	70	0.90	963.1271	918.9081	669.5515
		0.95	977.4810	930.1729	679.0749
		0.99	1004.8349	952.1849	697.1807

Table 3.5 – Values of the measure Δ for the approximating distributions for the l.r.t. 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}

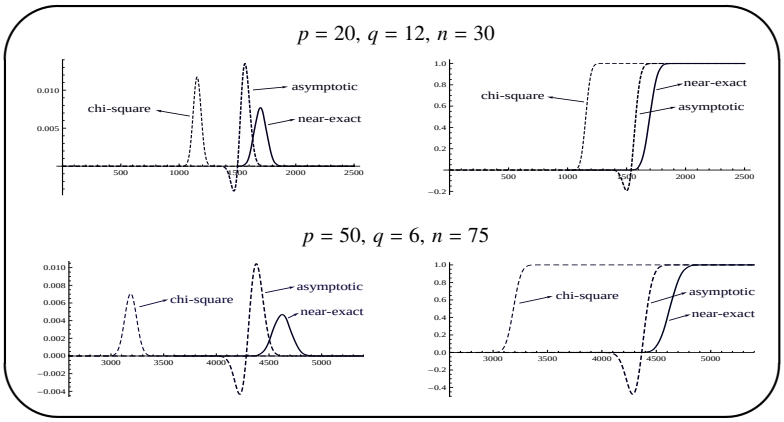

Figure 3.3 – Plots of p.d.f.'s and c.d.f.'s for approximating distributions of the l.r.t. statistic to test sphericity of the covariance matrix

Table 3.6 – Quantiles α for the approximating distributions for the l.r.t. statistic to test equality of q covariance matrices

p	q	n	α	Near-exact	Asymp.	Chi-square
20	12	30	0.90	1764.3372	1624.4478	1198.7624
			0.95	1783.7457	1639.0841	1211.4635
			0.99	1820.5701	1667.7083	1235.5291
	12	500	0.90	1219.3206	1216.8441	1198.7624
			0.95	1232.2401	1229.7374	1211.4635
			0.99	1256.7196	1254.1670	1235.5291
50	6	75	0.90	4732.0877	4486.2277	3280.9301
			0.95	4763.3410	4509.2761	3280.9301
			0.99	4822.3646	4554.8666	3320.3096

4. Conclusions

In short, near-exact distributions are asymptotic distributions built using a different concept, which leaves unchanged a good part of the original distribution, identified with a known manageable distribution, and approximates asymptotically the remaining part, in such a way that the whole corresponds to a known manageable distribution. This goal is usually achieved through an adequate factorization of the c.f. of the logarithm of the statistic under study. Near-exact distributions besides exhibiting very good performances for very small sample sizes, and an asymptotic behavior for increasing sample sizes, opposite to common asymptotic distributions, they also exhibit very good asymptotic behaviors for increasing numbers of variables involved, which is a very much welcome feature in Multivariate Analysis. In what concerns the answer to the question "why, or when, do we need them?", the answer is "when exact distributions are too complicated and common asymptotic distributions do not perform well", as it is often the case with most common asymptotic distributions for most of the l.r.t. statistics used in Multivariate Analysis, a fact completely overlooked by other authors.

Acknowledgements

This research was financially supported by CMA/FCT/UNL, under project PEst-OE/MAT/ UI0297/2011.

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.

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. Multivariate Anal.*, 89, 191–218.

Coelho, C. A., Marques, F. J. (2009). "The advantage of decomposing elaborate hypotheses on covariance matrices into conditionally independent hypotheses in building near-exact distributions for the test statistics", *Linear Algebra Appl.*, 430, 2592–2606.

Coelho, C. A., Marques, F. J. (2010). "Near-exact distributions for the independence and sphericity likelihood ratio test statistics", *J. Multivariate Anal.*, 101, 583–593.

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., 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.

Hwang, H.-K. (1998) "On convergence rates in the central limit theorems for combinatorial structures", *European J. Combin.*, 19, 329–343.

Loève, M. (1977) *Probability Theory*, vol. I, 4th ed., Springer, New York.