Combining p-Values and Random p-Values

M.F. Brilhante
Universidade dos Açores (DM), CEAUL
Rua da Mãe de Deus, Apartado 1422, 9501-801 Ponta Delgada, Portugal
E-mail: fbrilhante@uac.pt

D. Pestana
Universidade de Lisboa, Faculdade de Ciências (DEIO), CEAUL
Bloco C6 – Piso 4, Campo Grande, 1749-016 Lisboa, Portugal
E-mail: dinis.pestana@fc.ul.pt

F. Sequeira
Universidade de Lisboa, Faculdade de Ciências (DEIO), CEAUL
Bloco C6 – Piso 4, Campo Grande, 1749-016 Lisboa, Portugal
E-mail: fjsequeira@fc.ul.pt

Abstract. *Combining independent tests is a key-problem in Meta Analysis and one way of doing it is to use the reported p-values, which can be regarded as independent standard uniform observations under the common null hypothesis. Therefore, the problem of combining independent tests is closely tied to testing the uniformity of the p-values.*
Continuing the investigation initiated by Gomes et al. (2009), we present two simulation studies. In the first one we study the behaviour of pseudo-p-values, augmenting the sample of p-values using an auxiliary $Beta(1,q)$, $q \in [0.5, 3]$, *population;* $q = 3$, *namely, favors the generation of values close to zero, needed for an overall rejection of the null hypothesis. In the second case we study the behaviour of pseudo-p-values when these are internally generated, using an extension of the family of densities* f_{X_m}, $m \in [-2, 0)$, *introduced in Gomes et al. (2009), so that slopes* $m \in [0, 2]$ *are included.*
The results of the first study are by far better than those reported in Gomes et al. (2009), the proportion of rejections of the combined uniformity test increases with the number of pseudo-p-values. The results of the second study also show an increase of the power with data augmentation, specially when $m > -1$.

Keywords. Uniformity, combining p-values, random p-values, pseudo-p-values, data augmentation.

1. Introduction

Combining independent tests using observed p-values $(p_1, \ldots, p_n)$ to test a common null hypothesis H_0 *vs.* H_1 is an important problem in Meta Analysis. Under the common hypothesis H_0, as a direct result of the probability integral transform, we can consider that the p-values are independent observations of a random variable $P \frown \text{Uniform}(0,1)$. Hence, the problem of testing H_0 can simply be regarded as testing the uniformity of the sample of p-values (cf. Hartung *et al.*, 2008; Pestana, 2010). Tippett (1931) assessed the uniformity of the random sample testing whether its observed minimum $p_{1:n}$ could be an observation of $P_{1:n} \underset{|\,H_0}{\frown} Beta(1,n)$. This can of course be improved using more order statistics, but due to its simplicity Tippett's method is still in use. Fisher (1932), in the 4th edition of his path-breaking *Statistical Methods for Research Workers*, observed that as $-2\ln U \frown \chi^2_2$ when $U \frown Uniform(0,1)$, uniformity of the p-values can globally be tested using the fact that $-2\sum_{k=1}^{n} \ln p_k \frown \chi^2_{2n}$. Note that Tippett's method rejects the common null hypothesis H_0 at level α if $p_{1:n} < 1 - (1-\alpha)^{1/n}$, and Fisher's method if $-2\sum_{i=1}^{n} \ln p_i > \chi^2_{2n;\,1-\alpha}$. Pestana (2010) presents other ways of using directly p-values (as in Tippett's method) or functions of the

sample of p-values (as in Fisher's method) for combining the available p-values in order to reach an overall decision.

Although Birnbaum (1954) has shown that every monotone combined test procedure is *admissible*, *i.e.* provides a most powerful test against some alternative hypothesis for combining some collection of tests, and is therefore optimal for some combined testing situation, whose goal is to harmonize eventually conflicting evidence, or to pool inconclusive evidence, Littel and Folks (1971, 1973) have shown that under mild conditions Fisher's method is optimal for combining independent tests. Our findings confirm the superiority of Fisher's method.

There are two important aspects that need to be considered. On one hand, the number of reported p-values is usually small, and that affects the power of a combined uniformity test. On the other hand, publication bias implies that reported p-values are in general small (typically $p < 0.05$), and this is an important concern in Meta Analysis.

Computational intensive techniques such as the bootstrap have been used with success to deal with inference using small samples (Manly, 2007). Recently, data augmentation — cf. *v.g.* Royle *et al.* (2007) — has been used with arguable success, as Gomes *et al.* (2009) demonstrate: augmented samples can perform worse than the original small sample. This will be further discussed in Section 2, and the main focus of our research is to investigate ways of dealing with data augmentation that in fact do increase power in testing uniformity.

In what concerns the second problem, our investigation uses left-skewed auxiliary variables with support $[0, 1]$, more prone to generate values nearer 0 with high probability. Observe however that a side effect of the much discussed issue of publication bias in Meta Analysis is the eventual inappropriateness of simulation studies in the investigation of combining p-values; moreover, abundance of small p-values seems to point out that we are confronted with the much harder problem of random p-values (Kulinskaya *et al.*, 2008, p. 117–119) or of generalized p-values (Hartung *et al.*, 2008, p. 81–84), which are non-uniform.

2. Further issues on data augmentation

Gomes *et al.* (2009) applied the results of the following lemma (attributed in Johnson *et al.*, 1995, p. 285, to Deng *et al.*, 1992) to augment the sample of p-values, in order to get a larger set of the termed pseudo-p-values, hoping that the power of the test would increase.

LEMMA 1. *Let U and X be independent standard uniform random variables. Then*

$$W = \min\left(\tfrac{U}{X}, \tfrac{1-U}{1-X}\right)$$

and

$$V = X + U - \mathcal{I}[X + U]$$

are both standard uniform variables, with W and X independent and, V, X and U independent ($\mathcal{I}[x]$ denotes the largest integer not greater than x). In fact, the conclusion holds true, more generally, if X is an absolutely continuous random variable with support $(0, 1)$.

Gomes *et al.* (2009) also considered the family of density functions

$$f_{X_m}(x) = \left(mx - \tfrac{m-2}{2}\right) I_{(0,1)}(x)\,, \qquad (1)$$

$m \in [-2, 0)$, as an alternative to uniformity. Observe that, for $m \in [-2, 0)$, X_m is a mixture of a $Beta(1, 2)$ and a uniform random variables with mixing weights $-\frac{m}{2}$ and $1 + \frac{m}{2}$, respectively. As $X_m \prec X_0 \frown Uniform(0, 1)$ for all $m \in [-2, 0)$, members of this family are prone to generate values nearer 0, *i.e.* values one would expect for an overall rejection of the null hypothesis of uniformity. Obviously, if $m \in (0, 2]$, X_m is a mixture of the standard uniform and of a $Beta(2, 1)$, with weights $1 - \frac{m}{2}$ and $\frac{m}{2}$, respectively.

Observe that the inspiration to use this family comes from the observation that all segments with slope $m \in [-2, 2]$ passing through $(\frac{1}{2}, 1)$ are probability density functions of some variable X_m with support $(0, 1)$.

The proportion of rejections of uniformity was determined by simulation for sets

of pseudo-p-values when the initial sample $(p_1, \ldots, p_n)$ was generated from a population with density (1). What those authors observed was that the power of the test decreased with the number of pseudo-p's with both Tippett and Fisher's methods (these were the only methods compared).

These unexpected results were explained later by Sequeira (2009) and are due to a generalization of the previous lemma:

LEMMA 2. *Let X_{m_1} and X_{m_2}, m_1, $m_2 \in [-2, 0]$ be independent random variables from the family* (1). *Then*

$$W_{m_1,m_2} = \min\left(\frac{X_{m_1}}{X_{m_2}}, \frac{1-X_{m_1}}{1-X_{m_2}}\right) \stackrel{\mathrm{d}}{=} X_{\frac{m_1 m_2}{6}}.$$

Therefore, when using auxiliary uniform random variables, Gomes *et al.* (2009) were adding more uniform observations to the computationally augmented sample, contributing to smooth down the features that contradicted uniformity. Observe, more generally, that the uniform component is heavier in the resulting mixture random variable $X_{\frac{m_1 m_2}{6}}$ than in any of the initial X_{m_1} and X_{m_2} variables.

In the first simulation study, in the present work, we use instead independent $X \frown Beta(1, q)$ and $Y \frown Uniform(0, 1)$ (*i.e.* $Beta(1, 1)$) random variables in

$$\widetilde{W} = \min\left(\frac{X}{Y}, \frac{1-X}{1-Y}\right).$$

Analytical results are cumbersome, the general expression for the probability density function (for the even more general case of independent $Beta(p_\ell, q_\ell)$, $\ell = 1, 2$, random variables) being:

$$F_{\widetilde{W}}(w) = \frac{w^{p_2} \sum_{k=0}^{q_2-1} \frac{B(p_1+p_2+k, q_1)\binom{q_2-1}{k}(-w)^k}{k+p_2}}{B(p_1, q_1)\, B(p_2, q_2)} +$$

$$+ \frac{w^{q_2} \sum_{k=0}^{p_2-1} \frac{B(p_1, q_1+q_2+k)\binom{p_2-1}{k}(-w)^k}{k+q_2}}{B(p_1, q_1)\, B(p_2, q_2)}.$$

Our approach to the problem was to study, on one hand, the behaviour of the sample of p-values $\boldsymbol{p_n} = (p_1, \ldots, p_n)$ and the computationally augmented samples of pseudo-p-values

$$\boldsymbol{p_{2n}} = (p_1, \ldots, p_n, w_1^*, \ldots, w_n^*)$$

and

$$\boldsymbol{p_{3n}} = (p_1, \ldots, p_n, w_1^*, \ldots, w_n^*, v_1, \ldots, v_n),$$

when $\boldsymbol{p_n}$ is from a Beta$(1, q)$, $q \in [0.5, 3]$, distribution. In this case we have

$$w_k^* = \min\left(\frac{p_k}{u_k}, \frac{1-p_k}{1-u_k}\right),$$

$$v_k = p_k + w_k^* - \mathcal{I}[p_k + w_k^*],$$

$k = 1, \ldots, n$, and $u_1, \ldots, u_n$ independent pseudo-random numbers. As mentioned earlier we shall be particularly interested in the case $q = 3$ because this model favors the generation of p-values closer to zero.

The simulation results are presented in Section 3 and, contrarily to Gomes *et al.* (2009), we shall see that the simulated power of the combined uniformity test increases, in general, with the number of pseudo-p-values. An important remark: when using $Beta(1, q)$ auxiliary variables, $q \neq 1$, to generate the $\widetilde{W}$, we no longer have independence, although the dependence is not strong.

This is troublesome to establish analytical results, but has mild consequences in the simulation study. In fact, the more important Fisher's method relies on a chi-square test, and from the early stages of development of robustness there has been ample evidence that the lowering of degrees of freedom due to dependence issues in chi-square test is in general small (an extreme option is to use $0.7(n-1)$ in case the number of degrees of freedom under independence is n, cf. Mosteller and Tukey, p. 209), and henceforth that the chi-square test is in general liberal. On the other hand, dependence issues will influence the minimum of the sample, that will in principle exceed the minimum of an independent sample of the same size, from a fixed population, and thus Tippett's method will tend to be conservative. For example, in our particular case, the exact Tippett's critical point of $\min(X_1, \ldots, X_n, \widetilde{W}_1, \ldots, \widetilde{W}_n)$ when $X_k \frown Uniform(0, 1)$, and X_k and $\widetilde{W}_k$ are not independent, $k = 1, \ldots, n$, is $c^* = \frac{2}{3}[1 - (1-\alpha)^{1/n}]$ (note that the vectors $(X_k, \widetilde{W}_k)$, $k = 1, \ldots, n$, are independent). The value c^* differs from the used critical point $c = 1 - (1-\alpha)^{1/(2n)}$ when independence is assumed. However, the differences between the two points are quite small as one can see from the following table ($\alpha = 0.05$).

n	4	5	6	10	20
c^*	0.0085	0.0068	0.0057	0.0034	0.0017
c	0.0064	0.0051	0.0043	0.0026	0.0013

The exact critical point of $\min(X_1, \ldots, X_n, \widetilde{W}_1, \ldots, \widetilde{W}_n, V_1, \ldots, V_n)$, where $V_k = X_k + \widetilde{W}_k - \mathcal{I}[X_k + \widetilde{W}_k]$, in a dependence scenario is much more difficult to obtain.

Observe however that the question of validity (Hayes, 2005, p. 179: "it isn't totally accurate to talk about validity as a property of the test."). See further comments on the remarks on the simulation results, in Section 3.

We also decided to study the behaviour of the pseudo-p-values when these are "internally" generated, *i.e.*, when we randomly choose and fix one p-value, say p_j, to generate $w_k^* = \min\left(\frac{p_j}{p_k}, \frac{1-p_j}{1-p_k}\right)$, $k \neq j$. In this case we decided to use an extended version of the family of densities (1) with $m \in [-2, 2]$. The simulation results, also presented in section 3, will show that this strategy is an interesting one, specially when $m \in (-1, 0)$ (approximately).

3. Simulation results

For our first study we considered that the random p-value $P \frown \text{Beta}(1, q)$, $q \in [0.5, 3]$, and our objective was to determine the simulated power of the test $H_0 : q = 1$ (uniformity) *vs.* $H_1 : q \in [0.5, 3] \backslash \{1\}$. The outline for our simulation is as follows:

1. we generated 10,000 runs of samples of size $n = 4, 5, 6, 10, 20$ from a Beta$(1, q)$ population, $q = 0.5(0.05)3$, in order to determine the proportion of rejections of uniformity at levels 0.05 and 0.01 using Tippett and Fisher's methods;
2. we also defined
   (a) π_1, π_3 and π_5 as the proportion of rejections of uniformity using Tippett's rejection criterion with the samples $\boldsymbol{p_n}$, $\boldsymbol{p_{2n}}$ and $\boldsymbol{p_{3n}}$, respectively;
   (b) and π_2, π_4 and π_6 the proportion of rejections of uniformity using Fisher's method with the samples $\boldsymbol{p_n}$, $\boldsymbol{p_{2n}}$ and $\boldsymbol{p_{3n}}$, respectively.

Since $P_{1:n} \frown \text{Beta}(1, qn)$, it is straightforward to obtain the exact power function π_1. We have

$$\pi_1 = \pi_1(\alpha) = \Pr\left[P_{1:n} < 1 - (1-\alpha)^{1/n} \mid q\right] = 1 - (1-\alpha)^q .$$

We only show the results for $\alpha = 0.05$ because the pattern observed for other low value significance levels is similar. In the following figures (Fig. 1 and Fig. 2) the black solid lines correspond to Tippett's method and the black dashed lines to Fisher's method. The thickness of the lines increases with the number of pseudo-p's.

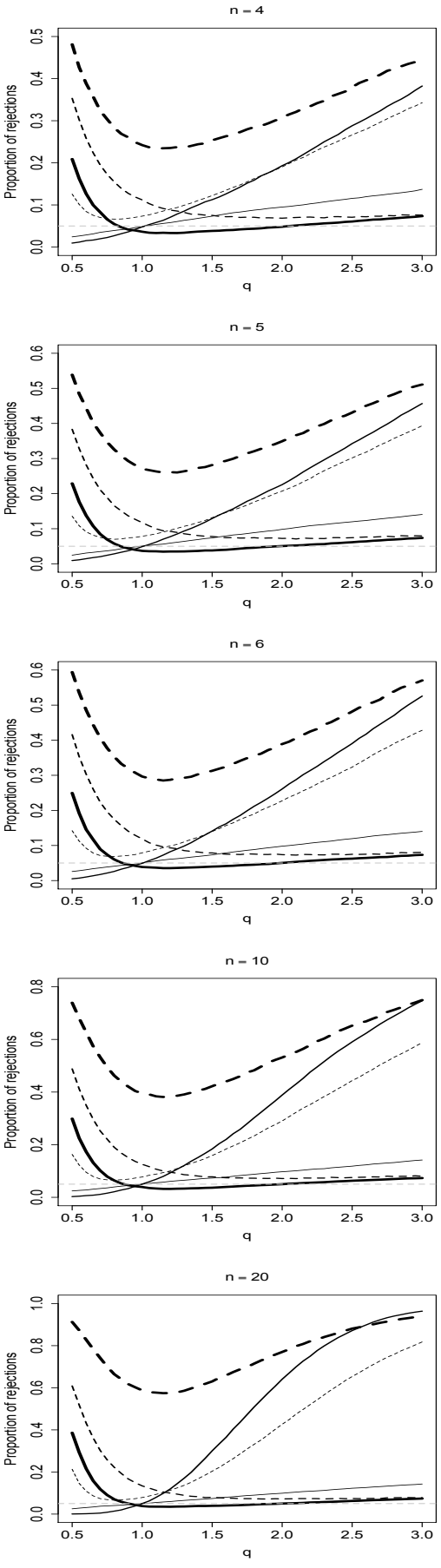

Fig. 1: Proportion of rejections, first simulation study.

Figure 1 shows that:

(i) Fisher's method performs better than Tippett's method in detecting departures from uniformity. Note that Fisher's rejection criterion has demonstrated in other simulation studies better performances than other p-values combining tests;

(ii) In general, $\pi_2 < \pi_4 < \pi_6$ (Fisher's method), indicating that the power increases with the number of pseudo-p-values;

(iii) For Tippett's method we see that $\pi_1 < \pi_3 < \pi_5$ for $q < 1$ (approximately) and $\pi_3 < \pi_5 < \pi_1$ for $q > 1$ (approximately);

(iv) When $q = 3$, the proportion of rejections of uniformity with the pseudo-p's p_{3n} is, overall, greater than 0.4 which can be considered a good result.

For our second study we considered an extension of the family of density functions (1) so that the slopes $m \in [0, 2]$ were included. Our strategy here was different from the first one since the pseudo-p-values were generated "internally". What we mean by this is that we randomly selected and fixed a p-value from the sample of p-values $(p_1, \ldots, p_n)$, say p_j, to generate

$$w_k^* = \min\left(\frac{p_j}{p_k}, \frac{1-p_j}{1-p_k}\right), \; k \neq j\,.$$

Our objective here was to determine the simulated power of the test $H_0 : m = 0$ (uniformity) *vs.* $H_1 : m \in [-2, 2]\backslash\{0\}$. The outline for this second simulation is as follows:

1. we generated 10,000 runs of samples of size $n = 4, 5, 6, 10, 20$ from a population X_m, $m = -2(0.05)2$, in order to determine the proportion of rejections of uniformity with Tippett and Fisher's methods;

2. we considered:

   (a) π_1 and π_3' the proportion of rejections of uniformity using Tippett's method with the samples $\boldsymbol{p_n} = (p_1, \ldots, p_n)$ and $\boldsymbol{p'_{2n-1}} = (p_1, \ldots, p_n, w_1^*, \ldots, w_{n-1}^*)$, respectively;

   (b) and π_2 and π_4' the proportion of rejections of uniformity using Fisher's criterion with the samples $\boldsymbol{p_n}$ and $\boldsymbol{p'_{2n-1}}$, respectively.

We did not study the simulated power of the set of pseudo-p's $\boldsymbol{p'_{3n-2}}$,

$$(p_1, \ldots, p_n, w_1^*, \ldots, w_{n-1}^*, v_1, \ldots, v_{n-1}),$$

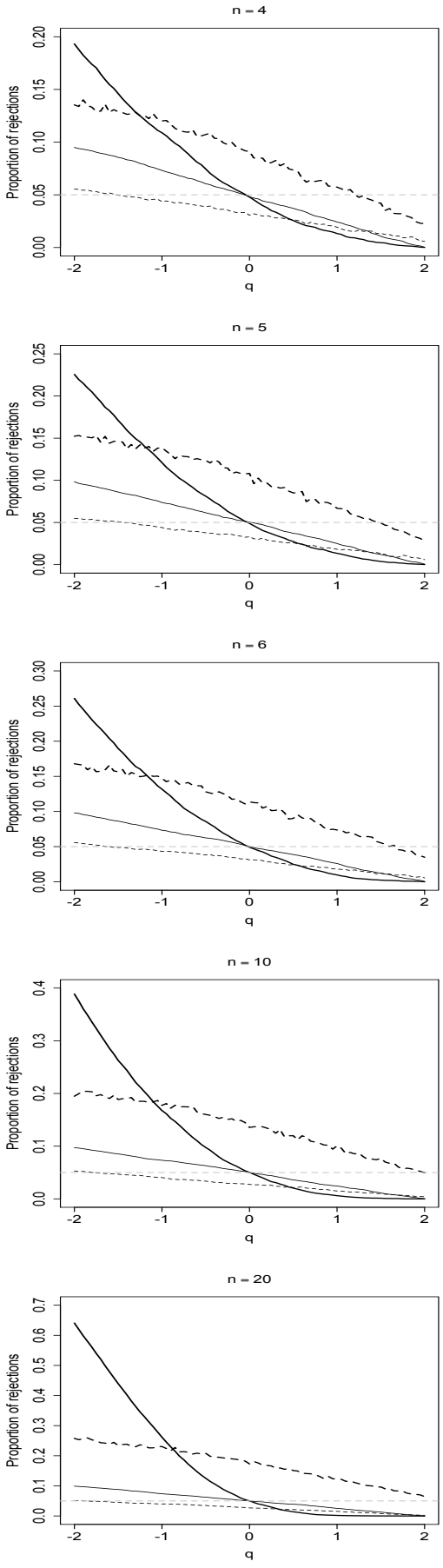

Fig. 2: Proportion of rejections, second simulation study.

where $v_k = p_k + w_k^* - \mathcal{I}[p_k + w_k^*]$, $k \neq j$, because it is expected a decrease in the proportion of rejections when compared to the proportion of rejections using $\boldsymbol{p'_{2n-1}}$.

For the same reasons as in the first case, we only show the results for $\alpha = 0.05$.

From Figure 2 we observe that:

(i) Again Tippett's method has the worst performance. The scenario here is even much worse than the one in the first study because, in general, the proportion of rejections is lower than the significance level;

(ii) For Fisher's method we have $\pi'_4 < \pi_2$ when $m < -1$, approximately, and $\pi_2 < \pi'_4$ if $m > 1$. However, the proportion of rejections can be lower than the significance level, specially when m approaches the value 2.

Quoting again Hayes (2005, p. 179): "So liberalness, conservativeness, and validity are properties not so much of tests themselves but the interaction between the test an the conditions in which the test is used."

4. Conclusions

Both strategies presented here for combining p-values are to a certain extent rewarding, specially when applying Fisher's criterion. With the first approach we obtained good results, in particular with the $Beta(1, 3)$ model which favors p-values closer to zero as expected under truthfulness of the overall null hypothesis, an expectation that is reinforced having in mind publication bias. The overall simulated power is greater than 0.4.

Although the second approach requires the determination of less pseudo-p-values, the simulated power is lower than in the first case (close to 0.15). Observe also that the best results are obtained when $-1 < m < 0$.

Acknowledgements

Research partially suported by FCT/OE.

The authors are grateful to the referees, whose comments pointed out weak points of the former draft, and had a substantial positive influence in improving the presentation.

References

[1] Birnbaum, A. (1954). Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49**, 559–575.

[2] Deng, L.-Y. and George, E.O. (1992). Some characterizations of the uniform distribution with applications to random number generation. *Ann. Instit. Statistical Mathematics*, **44**, 379–385.

[3] Fisher, R.A. (1932). *Statistical Methods for Research Workers*, 4th ed., Oliver and Boyd, London.

[4] Gomes, M.I., Pestana, D., Sequeira, F., Mendonça, S. and Velosa, S. (2009). Uniformity of offsprings from uniform and non-uniform parents, In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the ITI 2009, 31th International Conference on Information Technology Interfaces*, 243-248.

[5] Hartung, J., Knapp, G. and Sinha, B.K., (2008). *Statistical Meta-Analysis with Applications*, Wiley, New York.

[6] Hayes, A.F. (2005). *Statistical Methods for Communication Science*, Lawrence Erlbaum Associates, Mahwah, NJ.

[7] Johnson, N.L., Kotz, S., and Balakrishan, N. (1995). *Continuous Univariate Distributions*, vol. 2, 2nd ed., Wiley, New York.

[8] Kulinskaya, E., Morgenthaler, S., and Staudte, R. G. (2008). *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*, Wiley, Chichester.

[9] Littel, R. C., and Folks, L. J. (1971, 1973). Asymptotic optimality of Fisher's method of combining independent tests, I and II. *J. Amer. Statist. Assoc.* **66**, 802–806 and **68**, 193–194.

[10] Manly, B.F.J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 3rd ed., Chapman & Hall / CRC, Boca Raton.

[11] Mosteller, F., and Tukey, J.W. (1977). *Data Analysis and Regression*, Addison–Wesley, Reading, MA.

[12] Pestana, D. (2010). Combining p-values. In M. Lovric (ed.), International Encyclopedia of Statistical Sciences, Springer, ISBN: 978-3-642-04897-5

[13] Royle, J.A., Dorazio, R.M., and Link, W.A. (2007). Analysis of multinomial models with unknown index using data augmentation, *J. Computational Graph. Stat.*, **16**, 67–85.

[14] Sequeira, F. (2009). *Meta-Análise: Harmonização de Testes Usando os Valores de Prova*, Ph.D. Thesis, DEIO, Faculty of Sciences of the University of Lisbon.

[15] Tippett, L.H.C. (1931). *The Methods of Statistics*, Williams & Norgate, London.