Publication Bias and Meta-Analytic Syntheses

D. Pestana, M. L. Rocha, R. Vasconcelos and S. Velosa

Abstract Aside from more traditional methods of combining p-values, a test based on the geometric mean G_n of a uniform random sample of size n is developed. As $\mathbb{E}(G_n) = \left(\frac{n}{n+1}\right)^n \underset{n\to\infty}{\downarrow} 0.368$, it is obvious that publication bias has a bearing on the overall rejection of the null hypothesis, and that the recent concepts of random and of generalized p-values deserve full attention.

1 Introduction

Meta Analysis is a successful development of former systematic reviews, and is nowadays considered the goldstandard of reporting the previous findings by other researchers in Medicine, cf. the collection of invited papers by Egger and his co-authors published in the *British Medical Journal* (Egger and Davey Smith, 1997, 1997a, 1997b, 1978; Egger *et al.*, 1998; Davey Smith and Egger, 1998), Epidemiology (Woodward, 2005) and Pharmacology (Senn, 2007). The original development has been made by Glass (1976, 1978), an expert in Education Sciences, cf. also his

D. Pestana
Universidade de Lisboa, Faculdade de Ciências (DEIO) and CEAUL, Bloco C6, Piso 4, Campo Grande, 1749-016 Lisboa, Portugal, and CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa, e-mail: dinis.pestana@fc.ul.pt;fjsequeira@fc.ul.pt

M.. L. Rocha
Universidade dos Açores (DM) and CEAUL, Rua da Mãe de Deus, Apartado 1422, 9501-801 Ponta Delgada, Portugal, and CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa, e-mail: fbrilhante@uac.pt

R. Vasconcelos and S. Velosa
Universidade da Madeira (CCEE) and CEAUL, Campus Universitário da Penteada, 9000-390 Funchal, Portugal, and CEAUL, Centro de Estatística e Aplicações da Universidade de Lisboa, e-mail: smfm@uma.pt;sfilipe@uma.pt

interesting overview in Glass (1999). Recent developments appear in Hartung et al. (2008), Kulinskaya et al. (2008) and Borenstein et al. (2009).

Meta-analysis can be used to build up evidence from several inconclusive studies (namely because sample size is small and thus the power of tests is scarce); or to resolve conflicting evidence, when different studies, eventually concucted with different methodologies, seem to provide antagonistic results. A recent development of meta-analysis, christened cumulative meta-analysis, builds up evidence from costly and eventualy etically chalenging studies to draw the line when pooled significant results have been achieved.

Important journal in the area of Medicine, such as the British Medical Journal, nowadays recomend that substantial papers present a meta analysis of former results. This is possible because the publishing standards of research in Medicine attained some form of standardization in the presentation of evidence, that requires that statistical evidence is clearly reported means and standard deviations, observed values of the test statististics, or at least observed p-values. Under those circunstances proper meta-analysis can be performed, either presenting a global estimate of some measured effect, or combining p-values to achieve a global decision on some null hypothesis. This is so even when the studies have been conducted with very different precisions (a technique based on funnel plots provides in general interesting evidence), and even when very different treatments are compared, as for instance the celebrated studies on pre-eclampsia of pregnant women, where different treatments have been globally compared with a baseline diuretic.

Combining p-values is an important methods in meta-analysis (Pestana, 2009), since in most systematic reviews the only common reported statistical findings are p-values of tests on the same issue. The rationale is as follows: Let us assume that the p-values p_k are known for testing H_{0k} vs. H_{Ak}, $k = 1,\ldots,n$, in n independent studies on some common issue, and our aim is to achieve a decision on the overall question H_0^* : all the H_{0k} are true *vs*. H_A^* : some of the the H_{Ak} are true. As there are many different ways in which H_0^* can be false, selecting an appropriate test is in general unfeasible. On the other hand, combining the available p_k's so that $T(p_1,\ldots,p_n)$ is the observed value of a random variable whose sampling distribution under H_0^* is known is a simple issue, since under H_0^*, p is the observed value of a random sample $P = (P_1,\ldots P_n)$ from a $Uniform(0,1)$ population.

In what follows we describe methods that deal directly with the p-values (Tippett, Wilkinson, arithmetic mean), and methods that use transformed p-values (Fisher, Stoufer, logistic).

We also derive a new method using directly the p-values, using the fact that the probability density function and the distribution function of the geometric mean of an uniform random samples are easily obtained, and that the expected value and variance can easily be computed using the independence of the random variables. This is used to enhance the problem of publication bias.

2 Methods of combining p-values

A rational combined procedure should of course be *monotone*, in the sense that if one set of p-values $p = (p_1, \ldots, p_n)$ leads to rejection of the overall null hypothesis H_0^*, any set of componentwise smaller p-values $p' = (p'_1, \ldots, p'_n)$, $p'_k \leq p_k$, $k = 1, \ldots, n$, must also reject H_0^*.

Tippett (1931) used the fact that $P_{1:n} = \min\{P_1, \ldots, P_n\} \underset{|_{H_0^*}}{\frown} Beta(1,n)$ to reject H_0^* if the minimum observed p-value $p_{1:n} < 1-(1-\alpha)^{1/n}$. This *Tippett's minimum method* is a special case of *Wilkinson's method* (Wilkinson, 1951), advising rejection of H_0^* when some low rank order statistic $p_{k:n} < c$; as $P_{k:n} \frown Beta(k, n+1-k)$, to reject H_0^* at level α the cut-of-point c is the solution of $\int_0^c u^{k-1}(1-u)^{n-k}\mathrm{d}u = \alpha B(k, n+1-k)$.

Another way of using directly the observed p-values is to compute their arithmetic mean. However the exact distribution of $\overline{P}_n = \frac{1}{n}\sum_{k=1}^n P_k$ is cumbersome. For large n an approximation based on the central limit theorem can be used to perform an overall test on H_0^* vs. H_A^*, but as in general only a few p-values are available, this is the least used method of combining p-values.

In Section 2 we use the geometric mean

$$G_n = \sqrt[n]{\prod_{k=1}^n P_k}$$

of n independent uniform random variables, whose distribution function is readily computed, leading to a more powerful test based on the direct use of observed p-values; see, however, the discussion on publication bias in Section 4.

Alternatively, the construction of combined p-values using additive properties of simple functions of uniform random variables is a popular approach. Fisher (1932) used the fact that $P_k \frown Uniform(0,1) \Longrightarrow -2\ln(P_k) \frown \chi^2_2$, and therefore, $-2\sum_{k=1}^n \ln(P_k) \underset{|_{H_0^*}}{\frown} \chi^2_{2n}$. Then H_0^* is rejected at the significance level α if the $-2\sum_{k=1}^n \ln(p_k) > \chi^2_{2n,1-\alpha}$. Stouffer *et al.* (1949) used as test statistic $\sum_{k=1}^n \frac{\Phi^{-1}(P_k)}{\sqrt{n}} \underset{|_{H_0^*}}{\frown} Gaussian(0,1)$, where Φ^{-1} denotes the inverse of the distribution function of the standard gaussian, rejecting H_0^* at level α if $|\sum_{k=1}^n \frac{\Phi^{-1}(P_k)}{\sqrt{n}}| > z_{1-\alpha}$, where $z_{1-\alpha}$ stands for the $1-\alpha$ probability quantile of the standard gaussian.

Another simple transformation of uniform random variables P_k is the logit transformation, $\ln \frac{P_k}{1-P_k} \frown Logistic(0,1)$. As $\sum_{k=1}^{n} \frac{\ln\left(\frac{P_k}{1-P_k}\right)}{\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4)}}} \approx t_{5n+4}$, reject H_0^* at the significance level α if $-\sum_{k=1}^{n} \frac{\ln\left(\frac{p_k}{1-p_k}\right)}{\sqrt{n\,\frac{\pi^2(5n+2)}{3(5n+4}}} > t_{5n+4,\,1-\alpha}$.

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. In the context of social sciences Mosteller and Bush (1954) recommend Stouffer's method, but Littel and Folks (1971, 1973) have shown that under mild conditions Fisher's method is optimal for combining independent tests. Observe however that H_A^* states that some of the H_{Ak} are true, and so a meta-decision on H_0^* implicitly assumes that some of the P_k may have non-uniform distribution, cf. Hartung *et al.* (2008, p. 81–84) and Kulinskaya *et al.* (2008, p. 117–119), and references therein, on the promising concepts of generalized and of random p-values.

3 The geometric mean of a uniform random sample

The probability density function of n independent standard uniform random variables $U_1,\ldots,U_n$ is

$$f_{\pi_{k=1}^{n}}(x) = \frac{(-\ln(x))^{n-1}}{(n-1)!}\mathrm{I}_{[0,1)}(x)$$

and hence the probability density function of the geometric mean $G_n = (\prod_{k=1}^{n} P_k)^{\frac{1}{n}}$ of a random sample of size n from the standard uniform population is

$$f_{G_n}(x) = \frac{n\,[x(-n\ln(x))]^{n-1}}{\Gamma(n)}\mathrm{I}_{[0,1)}(x)$$

Fig. 1 below shows the probability density function of G_n for $n = 1,\ldots,20$ $\mathbb{E}\left(G_n^k\right) = \left(\frac{1}{1+\frac{k}{n}}\right)^n \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-k}$, and in particular $\mathbb{E}\left(G_n\right) = \left(\frac{n}{n+1}\right)^n \underset{n\to\infty}{\downarrow} \frac{1}{\mathrm{e}} \approx 0.3679$, the standard deviation decreases to zero, the skewness steadily decreases after a maximum 0.2645 for $n = 5$, and the kurtosis increases from -0.8541 ($n = 2$) towards 0.

In Table 1 we give the mean, variance, skewness and kurtosis of the geometric mean G_n of standard uniform random samples of size n, $n = 1,\ldots,20$.

The distribution function of G_n is

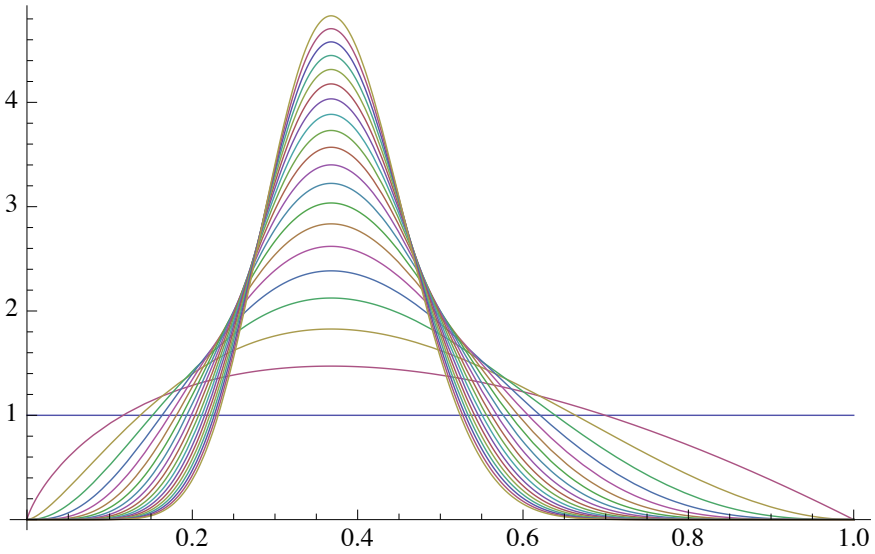

Fig. 1 Probability density functions of G_n, $n = 1, \ldots, 20$ (for $n = 1$, standard uniform; peakedness increases with n).

$$F_{G_n}(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ \dfrac{\Gamma^*(n, -\ln(x))}{\Gamma(n)} & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad 1 \le x \end{cases},$$

where $\Gamma^*(n,z)$ is the incomplete Gamma function $\Gamma^*(n,z) = \int_z^\infty x^{n-1} \mathrm{e}^{-x} \mathrm{d}x$.

The critical quantiles $g_{(n,1-\alpha)}$ such that $\int_0^{g_{(n,1-\alpha)}} \dfrac{(-\ln(x))^{n-1}}{(n-1)!} \mathrm{d}x = 1 - \alpha$ can be easily computed. In Table 2 we register the quantiles of probability $1 - \alpha$, $\alpha = 0.10,\ 0.05,\ 0.01$ of G_n.

4 Publication bias

The first step to carry over a meta-analysis is to select properly the evidence. In principle, a clear and fair criterion of inclusion must be adopted. Even so, bias publication must be taken into account, since non-significant results are rarely published in peer-reviewed journals (a general recommendation is to try to include properly chosen unpublished reports).

As in many other techniques used in meta-analysis, publication bias can easily lead to erroneous conclusions when combining p-values. In fact, most (if not all!)

Table 1 Mean value μ, variance σ^2, skewness γ_1, and kurtosis γ_2 of G_n, $n = 1,\ldots,20$.

n	μ	σ^2	γ_1	γ_2
1	0.5	0.0833333	0	-1.2
2	0.444444	0.0524691	0.18718	-0.854118
3	0.421875	0.0380215	0.24203	-0.640618
4	0.4096	0.0297587	0.260104	-0.505923
5	0.401878	0.0244288	0.264457	-0.415154
6	0.396569	0.0207112	0.262968	-0.350538
7	0.392696	0.0179723	0.258854	-0.30251
8	0.389744	0.0158715	0.25358	-0.26557
9	0.38742	0.0142095	0.247859	-0.236365
10	0.385543	0.012862	0.242051	-0.212747
11	0.383995	0.0117475	0.236342	-0.193284
12	0.382697	0.0108106	0.230825	-0.176991
13	0.381592	0.0100119	0.225543	-0.163163
14	0.38064	0.00932296	0.220513	-0.151291
15	0.379812	0.00872268	0.215736	-0.140993
16	0.379085	0.00819496	0.211205	-0.13198
17	0.378442	0.00772742	0.206908	-0.124029
18	0.377868	0.0073103	0.202834	-0.116965
19	0.377354	0.00693589	0.198968	-0.11065
20	0.376889	0.00659795	0.195296	-0.104971

available p-values come only from studies considered worth publishing because the observed p-values were small, seeming to point out significant results. Thus the assumption that the p_k's are observations from independent $Uniform(0,1)$ random variables is questionable, since in general they are in fact a set of low order statistics, given that p-values greater than 0.05, say, have not been recorded.

Observe, for instance, that whenever $p_{n:n}$ falls below the critical rejection point, the geometrical mean test studied in Section 3 will lead to the rejection of H_0^*, but $p_{n:n}$ smaller than the critical point (for $n \geq 14$, the expected value of G_n is greater than 0.36 and the standard deviation is smaller than 0.1) is what should be expected as a consequence of publication bias.

This obviously enhances one of the ill-resolved problems in meta-analysis: published results have in general significant values, typically less than 0.05. Hence most of the published studies point out that H_0 ought to be rejected, and that instead of combining p-values it would be more sensible to combine either generalized p-values or random p-values (Hartung it et al., 2008; Kulinskaya it et al., 2008).

A practical way of dealing with publication bias is to compute the number of umpublished studies with non-significant p-values that would be needed to reverse an overall decision of rejection of the null hypothesis, see Sequeira (2009) for details. A meta analysis on desmoplastic malignant melanoma, using the systematic review of Lens *et al.* (2005) and further evidence collected in Soares de Almeida *et al.* (2008), consultancy for researchers for other areas, and extensive simulation, namely with computationally augmented samples of p-values (Gomes *et al.*, 2008;

Table 2 Critical quantiles $1-\alpha$ of the geometric mean of uniform random samples..

$n \setminus \alpha$	0.10	0.05	0.01
2	0.143007	0.093300	0.036183
3	0.169635	0.122628	0.060690
4	0.188210	0.143932	0.081164
5	0.202156	0.160301	0.098183
6	0.213146	0.173397	0.112506
7	0.222110	0.184193	0.124741
8	0.229612	0.193300	0.135336
9	0.236016	0.201121	0.144623
10	0.241569	0.207937	0.152848
11	0.246448	0.213949	0.160200
12	0.250781	0.219305	0.166822
13	0.254663	0.224118	0.172830
14	0.258169	0.228475	0.178312
15	0.261357	0.232446	0.183341
16	0.264273	0.236083	0.187978
17	0.266953	0.239433	0.192271
18	0.269429	0.242531	0.196261
19	0.271726	0.245409	0.199982
20	0.273863	0.248090	0.203464
25	0.282708	0.259215	0.218040
30	0.289399	0.267661	0.229239
40	0.299025	0.279852	0.245586
50	0.305752	0.288396	0.257157
100	0.322999	0.310376	0.287301

Brilhante *et al.*, 2010), led to the conclusion that in what concerns decreasing power, and increasing number of unreported cases needed to reverse the overall conclusion of a meta-analysis, the methods of combining p-values rank as follows:

1. Arithmetic mean.
2. Geometric mean.
3. Chi-square transformation (Fisher's method).
4. Logistic transformation.
5. Gaussian transformation (Stouffer's method).
6. Selected order statistics (Wilkinson's method).
7. Minimum (Tippett's method).

References

1. Birnbaum, A.: Combining independent tests of significance. *J. Amer. Statist. Assoc.* **49**, 559–575 (1954).
2. Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R.: *Introduction to Meta-Analysis*, Wiley, Chichester (2009).

3. Brilhante, M. F., Mendonça, S., Pestana, D. and Sequeira, F.: Using Products and Powers of Products to Test Uniformity, submitted to the *32nd International Conference on Information Technology Interfaces*, (2010).
4. Brilhante, M. F., Pestana, D. and Sequeira, F.: Combining *p*-values and random *p*-values, submitted to the *32nd International Conference on Information Technology Interfaces*, (2010).
5. Davey Smith, G., and Egger, M.: Meta-Analysis — unresolved issues and future developments. *British Medical Journal* **316**, 221–225 (1998).
6. Egger, M., and Davey Smith, G.: Meta-Analysis — potentials and promise. *British Medical Journal* **315**, 1371–1374 (1997).
7. Egger, M., and Davey Smith, G.: Meta-Analysis — principles and procedures. *British Medical Journal* **315**, 1533–1537 (1997a).
8. Egger, M., and Davey Smith, G.: Meta-Analysis — beyond the grand mean? *British Medical Journal* **315**,1610–1614 (1997b).
9. Egger, M., and Davey Smith, G.: Meta-Analysis — bias in location and selection of studies. *British Medical Journal* **316**, 61–66 (1998).
10. Egger, M., Schneider, M., and Davey Smith, G.: Meta-Analysis — spurious precision? Meta-analysis of observational studies. *British Medical Journal* **316**, 140–144 (1998).
11. Fisher, R. A.: *Statistical Methods for Research Workers*, 4th ed., Oliver and Boyd, (1932).
12. Glass, G. V.: Primary, secondary, and meta-analysis of research. *Edu. Res.*, **5**, 3-8 (1976).
13. Glass, G.V.: Integrating findings: The meta-analysis of research. *Review of Research in Education*, **5**, 351-379 (1978).
14. Glass, G. V.: Meta-Analysis at 25, `http://glass.ed.asu.edu/gene/papers/meta25.html` (1999).
15. Gomes, M. I, Pestana, D., Sequeira, F., Mendonça, S., and Velosa, S.: Uniformity of offsprings from uniform and non-uniform parents. *Proceedings of the 31st International Conference on Information Technology Interfaces*, 243–248 (2009)
16. Hartung, J., Knapp, G., and Sinha, B. K.: *Statistical Meta-Analysis with Applications*, Wiley, New York (2008).
17. Kulinskaya, E., Morgenthaler, S., and Staudte, R. G.: *Meta Analysis. A Guide to Calibrating and Combining Statistical Evidence*, Wiley, Chichester (2008).
18. Lens, M. B., Newton-Bishop, J. A., and Boon, A. P.: Desmoplastic malignant melanoma: a systematic review. *British J. Dermatology*, **152**, 673–678 (2005).
19. Littel, R. C., and Folks, L. J.: Asymptotic optimality of Fisher's method of combining independent tests, I. *J. Amer. Statist. Assoc.* **66**, 802–806 (1971).
20. Littel, R. C., and Folks, L. J.: Asymptotic optimality of Fisher's method of combining independent tests, II. *J. Amer. Statist. Assoc.* **68**, 193–194 (1973).
21. Longford, N. T.: *Studying Human Populations* , Springer (2008).
22. Mosteller, F., and Bush, R.: Selected quantitative techniques, in G. Lidsey (Ed.), *Handbook of Social Psychology: Theory and Methods*, vol. I, Addison-Wesley, Cambridge, MA (1954).
23. Pestana, D.: Combining *p*-values. *Notas e Comunicações do CEAUL* (2009).
24. Senn, S.: *Statistical Issues in Drug Development*, 2nd ed., Wiley, Chichester (2007).
25. Sequeira, F.: *Meta-Análise: Harmonização de Testes Usando os Valores de Prova*. PhD Thesis, DEIO, Faculdade de Ciências da Universidade de Lisboa (2009).
26. Soares de Almeida, L., Requena, L., Rütten, A., Kutzner, H., Garbe, C., Pestana, D., and Marques Gomes, M.: Desmoplastic Malignant Melanoma: A Clinico-Pathologic Analysis of 113 Cases, *American J. Dermatopathology*, **30**, 207–215 (2008).
27. Stouffer, S. A., Schuman, E. A., DeVinney, L. C., Star, S. and Williams, R. M.: *The American Soldier*, vol. I: *Adjustment During Army Life*, Princeton University Press, Princeton (1949).
28. Tippett, L. H. C. : *The Methods of Statistics*, Williams & Norgate, London (1931).
29. Wilkinson, B.: A statistical consideration in psychological research. *Psychological Bulletin* **48**, 156–158 (1951).
30. Woodward, M.: *Epidemiology Study Design and Data Analysis*, 2nd ed., Chapman & Hall/CRC (2005).

Acknowledgements Research partially supported by FCT/OE.