Dealing with Generalized/Random p-Values

Maria de Fátima Brilhante, Dinis Pestana, and Fernando Sequeira

Abstract The use of generalised and of random p-values is mandatory when dealing with meta analysis of p-values to reach an overall conclusion on the result of independent tests on some issue. We explain the concepts, we indicate how to construct generalised test variables, nuisance parameters being or not being a concern, and present some examples. The Behrens-Fisher problem is revisited, providing an exact solution that avoids the fiducial confusion and that is less complex than the Welch-Satterthwaitte solution.

Keywords

Generalized and random p-values, combining p-values and uniformity

1 Introduction

Quite often decisions about a null hypotheses H_0 are made on the basis of p-values. The ordinary p-value is seen as a measure of evidence in favor of, or against, H_0, in the sense that small values give evidence that H_0 is false (or H_A is true), while big values give evidence of the opposite. However, when nuisance parameters are involved in testing problems, the p-value associated with a test statistic will usually depend on the nuisance parameters, which obviously is an undesirable feature. In order to overcome such difficulty, Tsui and Weerahandi [14] introduced the concept of generalized p-value, and by this eliminated the dependence of p-values on nuisance parameters when these are present.

Maria de Fátima Brilhante
Universidade dos Açores, Departamento de Matemática and
Universidade de Lisboa, CEAUL, Portugal
e-mail: fbrilhante@uac.pt

Dinis Pestana
Universidade de Lisboa, CEAUL and DEIO–FCUL,
Instituto de Investigação Científica Bento da Rocha Cabral, Portugal
e-mail: dinis.pestana@fc.ul.pt

Fernando Sequeira
Universidade de Lisboa, CEAUL and DEIO–FCUL
e-mail: fjsequeira@fc.ul.pt

Let X be a random variable whose distribution depends on the parameter of interest θ and the nuisance parameter η (real or vector), and let $X = (X_1, \ldots, X_n)$ be a random sample from the population X, where $x = (x_1, \ldots, x_n)$ is the observed value of X. In order to compute a generalized p-value for the testing problem $H_0 : \theta \leq \theta_0$ *vs.* $H_A : \ \theta > \theta_0$, for example, we first need the concept of a generalized test variable.

A random variable of the form $T = T(X; x, \theta, \eta)$ is a generalized test variable for the parameter θ if it satisfies the following properties: (i) the observed value of $T(X; x, \theta, \eta)$, *i.e.* $t = T(x; x, \theta, \eta)$, is free of θ and η; (ii) when θ is specified, the distribution of $T(X; x, \theta, \eta)$ is free of η; (iii) for fixed x and η, $\mathbb{P}(T \leq u; \theta)$ is a monotonic function of θ for any given u. Therefore, the generalized p-value is computed on the basis of a generalized test variable T as $p_G = \mathbb{P}(T \geq t | \theta = \theta_0)$, if $\Pr(T \leq u; \theta)$ is stochastically increasing with θ, with the assurance that p_G does not depend on η.

On the other hand, an important aspect of p-values, which is often forgotten, is its stochastic behavior. In practice, what is reported in scientific studies are the observed p-values. In fact, if F_0 denotes the distribution function of a test statistic T for θ under H_0, and if large values of T give evidence in favor of H_A being true, then the observed p value is $p \equiv p(x) = 1 - F_0(t)$. Thus, the random p-value associated with T is simply defined by $P \equiv P(X) = 1 - F_0(T)$, with P being a standard uniform random variable under H_0.

Now many meta-analytical syntheses involve the combination of reported p-values (e.g. in Pestana [8] various methods for combining p-values are indicated). The common thread amongst these methods is the fact that the considered p-values are, under the overall null hypothesis, a observed sample from a standard uniform distribution. Therefore, combining p-values methods and uniformity tests are closely related subjects. But as emphasized by some authors (e.g. Kulinskaya *et al.* [7]), when faced with a significant number of results that cast some doubt on the overall hypothesis, the correct approach to the problem should be to combine statistical evidence under the alternative hypothesis. We shall see how the concepts of generalized and random p-values can be used to deal with meta-analytical syntheses of evidence, specially in cases showing grounds to reject the overall hypothesis.

2 p-Values and their Role in Meta Analysis

The concept of p-value was introduced by Sir Ronald Fisher. The p-value is the probability, under the validity of H_0, of observing a value as extreme as or more extreme than the one observed for the test statistic, and therefore uses the data to assess the plausibility of a null hypothesis H_0.

In some statistics text books the p-value is defined as

- the lowest significance level α that forces rejection of H_0,
- and also the highest significance level α that allows for maintenance of H_0.

The p-value is a measure of evidence in favor of, or against, H_0, in the sense that:

- small values give evidence that H_0 is false, while
- big values give evidence of the opposite.

More formally, let $X = (X_1, \ldots, X_n)$ be a random sample from a population whose distribution depends on the parameter of interest θ.

A p-value $p(X)$ is a test statistic satisfying $0 \leq p(x) \leq 1$ for every sample point x.

A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$\mathbb{P}_\theta\big(p(X) \leq \alpha\big) \leq \alpha\,.$$

For the testing problem

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_A : \theta \in \Theta - \Theta_0$$

if $T = T(X)$ is a test statistic for θ, such that big values of T give evidence that H_A is true, then for each sample point x

$$p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}\big(T(X) \geq T(x)\big)\,.$$

In Meta-Analysis the p-values play an important role in combining independent tests.

Suppose that k independent tests where performed for the same testing problem, i.e. suppose we have

$$H_{0\mathrm{i}} \quad vs. \quad H_{A\mathrm{i}}\,, \quad i = 1, \ldots, k\,.$$

for which the p-value p_i, $i = 1, \ldots, n$, is known.

A combined test considers the global null hypothesis

$$H_0^* : \text{All } H_{0\mathrm{i}} \text{ are true} \quad vs. \quad H_A^* : \text{Some } H_{A\mathrm{i}} \text{ are true.}$$

Many methods for combining tests are based on p-values. All those methods are based on the fact that, under the validity of H_0, the observed p-values, $(p_1, \ldots, p_k)$, are a sample from a standard uniform population.

Therefore, the combination methods for p-values are based

- either on the properties of the uniform distribution,
- or on probability transformation methods.

2.1 Some methods for combining tests

- Tippett's method [13], or minimum P method — rejects H_0 at level α if

$$P_{1:k} < 1 - (1 - \alpha)^{1/k}$$

- Fisher's method [3] — rejects H_0 at level α if

$$-2 \sum_{i=1}^{k} \ln P_i > \chi^2_{2k;1-\alpha}$$

- Stouffer's method [12] — rejects H_0 at level α if

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} \Phi^{-1}(P_i) < -z_{1-\alpha}$$

- Geometric mean method [9], — rejects H_0 at level α if

$$\sqrt[k]{P_1 \cdots P_k} < G_{k;\alpha}$$

where $G_{k;\alpha}$ is the solution of the equation $x^k \sum_{j=0}^{k-1} \frac{(-k \ln x)^j}{j!} = \alpha$.

As an alternative to combining p-values, the combination of effect sizes from different studies is in most circumstances advisable.

Suppose that

- $\hat{\theta}_i$ is an estimate of the population effect size θ_i (e.g. $\theta_i = \mu_i$), obtained from population/study $i = 1, \ldots, k$, and
- we are interested in combining the effect sizes $\hat{\theta}_i$.

The combined estimate θ is given by the weighted combination of the $\hat{\theta}_i$'s, i.e.

$$\widetilde{\theta} = \frac{\sum_{i=1}^{k} \omega_i \hat{\theta}_i}{\sum_{i=1}^{k} \omega_i}$$

where the weights are reciprocal of variances $\omega_i = \dfrac{1}{\widehat{\mathrm{Var}}(\hat{\theta}_i)}$, $i = 1, \ldots, k$, as in most uses of weighted statistics.

Observe, however, that the pooling of statistical evidence from different studies makes sense only if the homogeneity hypothesis

$$H_0 : \theta_1 = \ldots = \theta_k = \theta \tag{1}$$

is not rejected.

There are many ways of verifying assumption (1), e.g. we can use Cochran's asymptotic chi-square test

$$X^2 = \sum_{i=1}^{k} \frac{(\hat{\theta}_i - \widetilde{\theta})^2}{\mathrm{Var}(\hat{\theta}_i)} .$$

The hypothesis (1) is rejected at level α if $X^2_{\mathrm{obs}} > \chi^2_{k-1;1-\alpha}$.

2.2 Homogeneity of means (homocedasticity) tests

Let X_{ij} be the j-th observation of the i-th population/study, such that

$$\mathbb{E}(X_{ij}) = \mu_i \quad \text{and} \quad \mathrm{Var}(X_{ij}) = \sigma_i^2$$

$i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

Some (approximate) tests for $H_0 : \mu_1 = \ldots = \mu_k$ are:

- Cochrans's test (standard test for testing homogeneity in Meta-Analysis)
- Welch-Satterthwaite test
- Brown-Forsythe test
- Mehrotra (Modified Brown-Forsythe) test
- Approximate ANOVA F test
- Ajusted Welch-Satterthwaite test

This multiplicity of solutions stems out from the fact that in this celebrated Behrens-Fisher problem it was observed, for the first time, that the frequentist and the bayesian methodologies could produce different results. Public criticism of Fishers' solution, made by Bartlett, that commented that the confidence coefficient announced didn't match the coverage probability of the regional estimator, irritated Fisher to the extent of leading him to create the rather mysterious concept of fiducial inference, to completely disregard Welch's approximate solution, and to accept Jeffrey's bayesian solution (coincidental with the one he had derived) in the *Annals of Eugenics*

For details, and comments on the revolutionary Scheffé's [10, 11] randomised paired test, cf. [2]. In Section 3 we present an example with an exact solution, easily extended to k comparisons, using generalised statistics, cf. also [14].

3 Generalized p-Values

Simulation studies have shown that the Modified Brown-Forsythe and the Approximate F tests are more robust for departures from normality with homogeneous variances.

When nuisance parameters are present, testing homogeneity can be somewhat challenging — the associated p-value will depend on the nuisance parameters.

However, exact solutions can be found by simply using the concept of generalized p-value introduced by Tsui and Weerahandi [14] in order to eliminate their dependence on nuisance parameters.

3.1 Generalized test variable and generalized p-value

A r.v. of the form $T = T(X;x,\theta,\eta)$ is a generalized test variable (g.t.v.) for the parameter θ if it satisfies the following properties:

1. The observed value of $T(X;x,\theta,\eta)$, i.e. $T(x;x,\theta,\eta)$, is free of θ and η;
2. When θ is specified, the distribution of $T(X;x,\theta,\eta)$ is free of η;
3. For fixed x and η, $\mathbb{P}(T \leq t;\theta)$ is a monotonic function of θ for any given t.

Let X be a r.v. whose distribution depends on the parameter of interest θ and the nuisance parameter η, and that we are interested in testing

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_A : \theta > \theta_0 \,. \tag{2}$$

If $T(X;x,\theta,\eta)$ is a g.t.v. stochastically increasing with θ, than the generalized p-value for problem (2) is computed as

$$p_G = \mathbb{P}(T(X;x,\theta,\eta) \geq T(x;x,\theta,\eta)|\theta = \theta_0).$$

3.2 The Behrens-Fisher problem — an exact solution

Let $X_1 = (X_{11},\ldots,X_{1n_1})$ and $X_2 = (X_{21},\ldots,X_{2n_2})$ be independent random samples from the populations $X_1 \frown N(\mu_1,\sigma_1)$ and $X_2 \frown N(\mu_2,\sigma_2)$, respectively.

We want to test $H_0 : \mu_1 = \mu_2$ *vs.* $H_A : \mu_1 \neq \mu_2$.

Let

- $X = (\overline{X}_1 - \overline{X}_2, S_1^2, S_2^2)$
- $x = (\bar{x}_1 - \bar{x}_2, s_1^2, s_2^2)$
- $\theta = \mu_1 - \mu_2 \quad \longrightarrow$ parameter of interest
- $\eta = (\sigma_1^2, \sigma_2^2) \quad \longrightarrow$ nuisance parameter

A generalised test variable for this problem is

$$T = T(X;x,\theta,\eta) = \frac{\overline{X}_1 - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sqrt{\frac{s_1^2\sigma_1^2}{S_1^2 n_1} + \frac{s_2^2\sigma_2^2}{S_2^2 n_2}} \tag{3}$$

where $T_{\text{obs}} = T(x;x,\theta,\eta) = \bar{x}_1 - \bar{x}_2$. and $\mathbb{E}(T)$ increases with $\theta = \mu_1 - \mu_2$.

The variable (3) can be rewritten as

$$T = Z\sqrt{\frac{(n_1-1)s_1^2}{n_1\, Y_1} + \frac{(n_2-1)s_2^2}{n_2\, Y_2}}$$

where $Z \frown N(0,1)$, $Y_1 \frown \chi^2_{n_1-1}$ and $Y_2 \frown \chi^2_{n_2-1}$ are independent r.v.'s.

Therefore,

$$\begin{aligned} p_G &= \mathbb{P}\left(\left| Z\sqrt{\frac{(n_1-1)s_1^2}{n_1 Y_1} + \frac{(n_2-1)s_2^2}{n_2 Y_2}} \right| \geq |\bar{x}_1 - \bar{x}_2| \,\middle|\, \theta = 0 \right) \\ &= \mathbb{P}\left(Z^2 \left(\frac{(n_1-1)s_1^2}{n_1 Y_1} + \frac{(n_2-1)s_2^2}{n_2 Y_2} \right) \geq (\bar{x}_1 - \bar{x}_2)^2 \right) \end{aligned}$$

that can be obtained by simulation.

Example 1: (Yield of corn from two hybrids)

Hybrid	n_i	$\bar{x}_i$	s_i
A	6	6.75	0.435
B	5	7.02	0.172

(Hartung et al., [5], p. 69)

- Generalized p-value

$$p_G = \mathbb{P}\left(Z^2 \left(\tfrac{0.158}{Y_1} + \tfrac{0.024}{Y_2} \right) \geq 0.0729 \right) = 0.238$$

(based on 5 000 runs)

- Cochran's test

$$\widetilde{\mu} = 6.977 \longrightarrow X^2_{\text{obs}} = 1.946 \longrightarrow p = 0.163$$

3.3 A useful recipe for obtaining generalized test variables

Suppose that the parameter of interest is $\theta = f(\theta_1,\ldots,\theta_k)$, where the θ_i's are unknown parameters. Following [6], a general methodology for constructing a generalised test statistic is as follows:

1. Find a set of (sufficient) statistics $(U_1,\ldots,U_k)$ and a set of invertible pivots $(V_1,\ldots,V_k)$ relating $(U_1,\ldots,U_k)$ to $(\theta_1,\ldots,\theta_k)$;
2. Express $\theta = g(U_1,\ldots,U_k,V_1,\ldots,V_k)$;
3. $T = \theta - g(u_1,\ldots,u_k,V_1,\ldots,V_k)$ is a generalized test variable, stochastically increasing with θ.

Example 2:
Let X_1 and X_2 be two independent random samples, respectively from the populations $X_1 \frown \text{Exp}(\lambda_1, \delta_1)$ and $X_2 \frown \text{Exp}(\lambda_2, \delta_2)$.

We are interested in testing

$$H_0 : \lambda_1 - \lambda_2 \leq \theta_0 \quad vs. \quad H_A : \lambda_1 - \lambda_2 > \theta_0 \,. \tag{4}$$

As

- $\theta = \lambda_1 - \lambda_2 \quad \longrightarrow$ parameter of interest
- $\eta = (\delta_1, \delta_2) \quad \longrightarrow$ nuisance parameter

$(\overline{X}_1, \overline{X}_2, X_{1;1:n_1}, X_{2;1:n_2})$ is a vector of sufficient statistics for $(\lambda_1, \lambda_2, \delta_1, \delta_2)$.

- $V_1 = \frac{X_{1;1:n_1} - \lambda_1}{\delta_1} \frown \text{Gamma}(1, n_1^{-1})$
- $V_2 = \frac{\overline{X}_1 - X_{1;1:n_1}}{\delta_1} \frown \text{Gamma}(n_1 - 1, n_1^{-1})$
- $V_3 = \frac{X_{2;1:n_2} - \lambda_2}{\delta_2} \frown \text{Gamma}(1, n_2^{-1})$
- $V_4 = \frac{\overline{X}_2 - X_{2;1:n_2}}{\delta_2} \frown \text{Gamma}(n_2 - 1, n_2^{-1})$,

solving with respect to the unknown parameters:

- $\lambda_1 = X_{1;1:n_1} - (\overline{X}_1 - X_{1;1:n_1}) \frac{V_1}{V_2}$
- $\delta_1 = \frac{\overline{X}_1 - X_{1;1:n_1}}{V_2}$
- $\lambda_2 = X_{2;1:n_2} - (\overline{X}_2 - X_{2;1:n_2}) \frac{V_3}{V_4}$
- $\delta_2 = \frac{\overline{X}_2 - X_{2;1:n_2}}{V_4}$.

Applying Hanning *et al.* [6] methodology to problem (4), we get

$$\begin{aligned} T &= \theta - (x_{1;1:n_1} - x_{2;1:n_2}) + (\bar{x}_1 - x_{1;1:n_1}) \frac{V_1}{V_2} - (\bar{x}_2 - x_{2;1:n_2}) \frac{V_3}{V_4} \\ &= \theta - (x_{1;1:n_1} - x_{2;1:n_2}) + \frac{(\bar{x}_1 - x_{1;1:n_1})}{n_1 - 1} \, Y_1 - \frac{(\bar{x}_2 - x_{2;1:n_2})}{n_2 - 1} \, Y_2 \end{aligned}$$

with $Y_1 \frown F_{2,2(n_1-1)}$ and $Y_2 \frown F_{2,2(n_2-1)}$ independent r.v.'s.

As $T_{\text{obs}} = 0$,

$$p_G = \mathbb{P}\Big(\frac{(\bar{x}_1 - x_{1;1:n_1})}{n_1 - 1} Y_1 - \frac{(\bar{x}_2 - x_{2;1:n_2})}{n_2 - 1} Y_2 \geq x_{1;1:n_1} - x_{2;1:n_2} - \theta_0 \Big) \,.$$

4 Random p-Values

We often forget that p-values have a stochastic behavior.

Let

- T be a continuous test statistic for testing hypotheses about θ;
- F_0 be the df of T under H_0;
- F_θ be the df of T under some alternative H_A.

If large values of T give evidence in favor of H_A, then the observed p-value is given by

$$p = \mathbb{P}(T \geq t|H_0) = 1 - F_0(t)\,.$$

Therefore, the random p-value associated with the test is the r.v.

$$P = 1 - F_0(T)$$

where $P \frown U(0,1)$ under H_0.

On the other hand, the df of P under H_A is given by

$$\begin{aligned} \mathbb{P}_\theta(P \leq p) &= \mathbb{P}_\theta(1 - F_0(T) \leq p) \\ &= 1 - F_\theta(F_0^{-1}(1-p)), \quad 0 < p < 1. \end{aligned} \tag{5}$$

4.1 Why should we bother with random p-values?

When faced with a significant number of results that cast some doubt on the global hypothesis, the correct approach should be to combine evidence under the alternative hypothesis.

Some authors recommend the use of the expected p-value, EPV, under H_A, as a measure of performance of a test — the smaller the value, the better the test.

Advantages:

- it depends on the alternative, not on the significance level;
- it allows to determine the sample size;
- it allows to determine which alternative the observed p-value represents.

Disadvantages:

- the distribution of P under H_A is usually highly skewed, and therefore the EPV is not a good measure to represent its distribution.

Example 3
Let us consider the family of pdf's

$$f_m(x) = \left(mx - \tfrac{m-2}{2}\right) \mathrm{I}_{(0,1)}(x)\,, \quad m \in [-2,0]$$

(Gomes et al., [4]), and the testing problem

$$H_0 : m = 0 \ \text{(uniformity)} \quad \textit{vs.} \quad H_A : m \in [-2,0)\,.$$

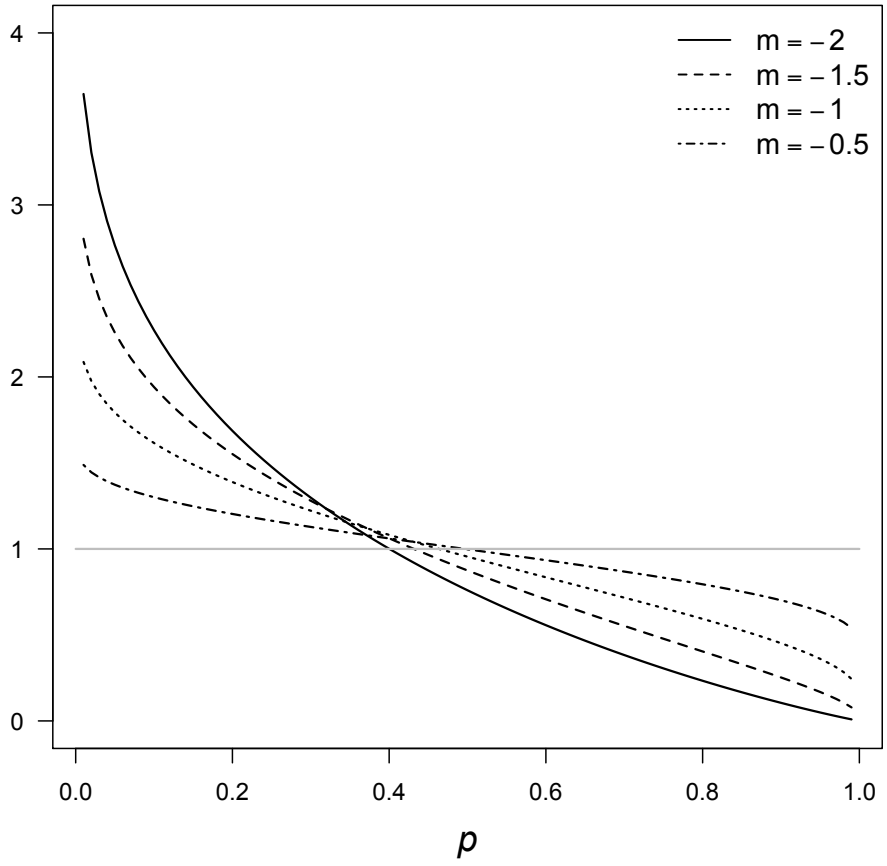
Let us also consider Fisher's test statistic for combining p-values, i.e.

$$T_n = -2\sum_{i=1}^{n} \ln P_i\,.$$

Under H_A, the random p-value P associated with Fisher's test has pdf

$$f_P(p) = \frac{f_m(\chi^2_{2n;1-p})}{f_0(\chi^2_{2n;1-p})}, \quad 0 < p < 1.$$

Fig. 1 Probability density function of P associated with T_3 under some alternatives

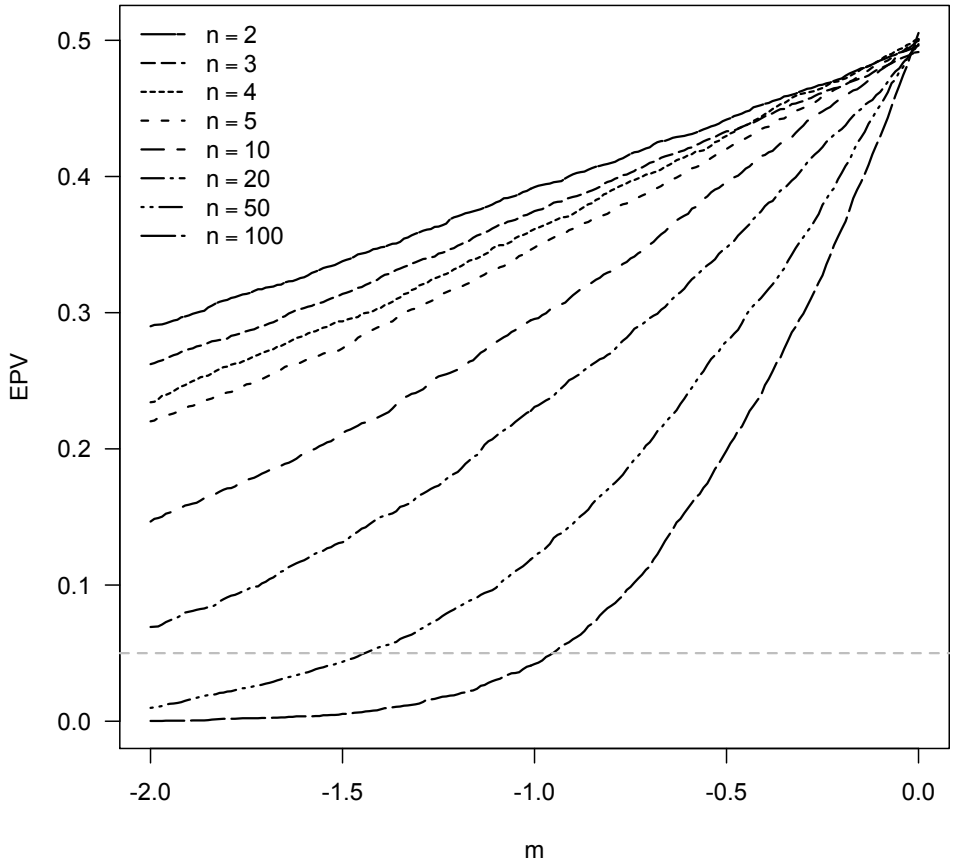


5 Conclusions

Generalised test statistics and the ensuing generalised p-values are a sharp instrument in inference when nuisance parameters preclude the use of traditional methods. In more general settings, those tools can improve accuracy or power.

Random p-values bring in a deeper understanding of inferential procedures, since statistical analysis must be followed by a critical appraisal of conclusions drawn from the statistical evidence. Moreover, testing

Fig. 2 EPV under the alternative hypothesis using Fisher's test statistic



overall composite hypothesis must always have in focus the consequences of a true alternative hypothesis in some of the experiments.

Acknowledgements

This research has been supported by National Funds through FCT — Fundação para a Ciência e a Tecnologia, project PEst-OE/MAT/UI0006/2011.

References

1. Brilhante, M.F. (2013). Generalized *p* Values and Random *p* Values when the Alternative to Uniformity is a Mixture of a Beta(1,2) and Uniform. In Oliveira, P. *et al.* (eds), *Recent Developments in Modeling and Applications in Statistics*, Springer, Heidelberg, 159-167.
2. Brilhante, M. F., Pestana, D. D., Rocha, J., Rocha, M. L., e Velosa, S. (2011). *Inferncia Estatstica Sobre a Localizao Usando a Escala*, Instituto Nacional de Estatstica, Lisboa.
t

3. Fisher, R. A. (1935). The fiducial argument on statistical inference. Ann. Eugen. 6, 391398.
4. Gomes, M. I, Pestana, D., Sequeira, F., Mendonça, S., and Velosa, S.: Uniformity of offsprings from uniform and non-uniform parents. In Luzar-Stiffler, V., Jarec, I. and Bekic, Z. (eds.), *Proceedings of the 31st International Conference on Information Technology Interfaces*, 243–248 (2009).
5. Hartung, J., Knapp, G. and Sinha, B. (2008). Statistical Meta-Analysis with Applications, Wiley, New Jersey.
6. Hanning, J., Iyer, H. K., and Patterson, P. D. (2006). Fiducial generalized confidence Intervalas, *J. Amer. Statist. Ass.* **101**, 254–269.
7. Kulinskaya, E., Morgenthaler, S. and Staudte, R.G. (2008). Meta Analysis: A Guide to Calibration and Combining Statistical Evidence, Wiley, Chichester.
8. Pestana, D. (2011). Combining p-values. In Lovric (ed.) International Encyclopedia of Statistical Science, 1145-1147, Springer, Berlin.
9. Pestana, D., Rocha, M. L., Vasconcelos, R., and Velosa, S. (2013). Publication Bias and Meta-Analytic Syntheses. In Lita da Silva, J.; Caeiro, F.; Natrio, I.; Braumann, C.A. (Eds.): *Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and other Statistical Applications*, 347-354, Springer, Berlin.
10. Scheffé, H. (1943). On solutions of the Behrens-Fisher problem based on the $t-$distribution. *Ann. Math. Stat.* **14**, 35–44.
11. Scheffé, H. (1944). A note on the Behrens-Fisher problem. *Ann. Math. Stat.* **15**, 430–432.
12. Stouffer, S. A., Schuman, E. A., DeVinney, L. C., Star, S., and Williams, R. M. (1949). *The American Soldier*, vol. I: *Adjustment During Army Life*, Princeton University Press, Princeton.
13. Tippett, L. H. C. : *The Methods of Statistics*, Williams & Norgate, London (1931).
14. Tsui, K. and Weerahandi, S. (1989). Generalized p-values in significance testing hypothesis. *J. Am. Stat. Association*, **84**, 602-607.