This is the peer reviewed version of the following article:

Cabras, S. & Castellanos, M. E. (2017). P-value calibration in multiple hypotheses testing. *Statistics in Medicine*, 36(18), pp. 2875–2886.

which has been published in final form at

<https://doi.org/10.1002/sim.7330>

This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for [Use of Self-Archived Versions](#).

P-value calibration in multiple hypotheses testing

Stefano Cabras^{a,b,*†} and Maria Eugenia Castellanos^c

As *p*-values are the most common measures of evidence against a hypothesis, their calibration with respect to null hypothesis conditional probability is important in order to match frequentist unconditional inference with the Bayesian ones. The Selke, Bayarri and Berger calibration is one of the most popular attempts to obtain such a calibration. This relies on the theoretical sampling null distribution of *p*-values, which is the well-known Uniform(0,1), but arising only for specific sampling models. We generalize this calibration by considering a sampling null distribution estimated from the data. It is possible to obtain such an empirical null distribution, for instance, in the context of multiple testing in which many *p*-values come from the null model. Such a context is purely instrumental for the purposes of *p*-value calibration, and multiple testing still needs to be considered with appropriate techniques. The new calibration proposed here still remains a simple analytic formula like the original one under the Uniform(0,1) and basically provides a stronger interpretation framework for the widely used *p*-value.

Keywords: Bayes factor lower bound; non-parametric Bayes; objective Bayes; significance intestine

1. Motivation and Introduction

P-values or tail areas are the most commonly used measures to quantify evidence against a hypothesis; however, their use in statistical testing leads to well-known drawbacks that also motivated the American Statistical Association to publish a statement on *p*-values [1]. Among these drawbacks, we concentrate on the calibration problem, which is motivated here with a simple example. Suppose testing many hypotheses with many corresponding *p*-values: some of them are less than, say 10^{-5} , and the rest greater. Suppose we decide to reject (or more appropriately put in question) those null hypotheses with $p < 10^{-5}$ and not question or not reject the rest of the hypotheses. Rejection for $p < 10^{-5}$ is based on grounds that these

p-values are surprising under the null, and this is because if under the null, the *p*-values are uniformly distributed, then observing $p < 10^{-5}$ should have a very small probability and the evidence of the null conditional to the data would still be very low. Suppose now that the uniformity assumption is questioned or cannot be assumed, then we do not know if $p < 10^{-5}$ has a very high or low probability under the null. In fact, unless another distribution assumption replaces the uniform one, it is not possible to know whether $p < 10^{-5}$ is surprising or not, or what the conditional probability of the null is. Proposing such an alternative assumption about the sampling null distribution and providing the corresponding conditional probability of each test is the aim of this work in the context of multiple hypotheses testing (MHT). Our purpose is not to deal with MHT nor to promote the *p*-value as the only meaningful decision tool but certainly to provide a stronger framework for its interpretation and also for considering other alternative probabilistic measures such as the conditional type I error further defined. These, along with a sensitivity analysis on the data, should guide the null hypotheses questioning (rejection) process.

We formalize the aforementioned concepts with mathematical symbols, trying to keep the notation as simple as possible. In many applications, data $\mathbf{X}$ is modeled through a null model, $H_0 : \mathbf{X} \sim f_0(\mathbf{x})$. The *p*-value is calculated with the purpose of choosing between this null model and a barely defined

^aDepartment of Statistics, Universidad Carlos III de Madrid, Getafe, Spain

^bDepartment of Mathematics and Informatics, Università di Cagliari, Cagliari Italy

^cDepartment of Informatics and Statistics, Universidad Rey Juan Carlos (Móstoles, Spain)

*Correspondence to: Stefano Cabras, Department of Statistics, Universidad Carlos III de Madrid, C/Madrid, 128 – 28903 Getafe, Spain.

†E-mail: s.cabras@unica.it

alternative one denoted by H_1 . Usually this is performed using a statistic $T(\mathbf{X})$, for which and without loss of generality we assume that large values of T indicate incompatibility of H_0 . Given this, the well-known p -value is $p = \Pr_{H_0}(T > t_{obs} = T(\mathbf{x}_{obs}))$, defined upon a suitable sampling distribution of T under H_0 . As t_{obs} can be observed under H_0 or H_1 , then the corresponding p could have been observed under H_0 or H_1 , then the hypothesis testing can be reformulated as deciding which is the most probable p -value generating mechanism given the observed data. Therefore, the choice among H_0 or H_1 is addressed here through a Bayesian model selection problem in the p -value scale rather than, as usual, in the scale of the observed $T(\mathbf{x}_{obs})$ test statistic.

The aforementioned setup is applied, for example, in microarray experiments, in which the comparison between gene expressions under two independent biological populations is performed using two sample t -tests summarized through p -values. Similar types of tests are used to analyze proteomic data, or imaging studies comparing activation areas in subjects under different conditions, (see examples in [2]), just to cite one of the most popular and massive applications of p -values. Such a massive use of tests leads to the well-known multiplicity problem, which then leads to the MHT approach that here is purely instrumental for our purposes of p -value calibration, as explained later.

In [3] and [4], the authors propose a calibration of the p -value that is easier to be interpreted. In particular, given a p -value, the Bayesian calibration is

$$\underline{B}(p) = -ep \log p \quad (1)$$

when $p < 1/e$ and this is interpreted as a lower bound on the odds provided by the data (or Bayes factor, BF) for H_0 to H_1 .

Result in (1) will be referred to as the Sellke–Bayarri–Berger (SBB) calibration of the p -value with respect to the BF lower bound [3]. The SBB calibration can also be expressed in terms of the lower bound of the null hypothesis posterior probability, $\underline{\alpha}(p)$, assuming equal prior probability for each hypothesis, $\pi(H_0) = \pi(H_1) = 1/2$, that is,

$$\underline{\alpha}(p) = 1/(1 + \underline{B}(p)). \quad (2)$$

The value $\underline{\alpha}(p)$ is also referred to as the lower bound of the frequentist type I error probability [3], and this achieves the matching between unconditional frequentist inference and the conditional Bayesian one. Intuitively, this is the minimum *a posteriori* probability that can be assigned to H_0 given the observed p . Subsection A.1 in the Appendix reports more technical details about this calibration.

The main assumption in SBB calibration, relaxed in the present work, is the p -value uniformity under the null hypothesis H_0 . This is, of course, the main reference distribution of the p -value under the null that will be referred to as the theoretical sampling null distribution of the p -value. Such a distribution can be assumed only in very few cases: either T is ancillary with respect to unknown nuisance parameters; that is, the p -value was calculated over a pivotal quantity, or the distribution of T under H_0 , $G_0(t)$, is fully known, that is the null distribution is a ‘precise’ one. Note that in some cases, the distribution of T is known only asymptotically with respect to the sample size n used to calculate t_{obs} , and for n finite the sampling null distribution of the p -value is practically unknown. As discussed in [5], there are many reasons that invalidate the theoretical null distribution, such as, for example, the presence of nuisance parameters, restrictive assumptions about data as independence, and asymptotic approximation. Calculation of $\underline{B}(p)$ is then reliable only for very few hypothesis testing situations, where such reliability will be clarified and measured later.

Consider the following real study about prostate cancer data [6], $m = 6033$ genes comparing $n_x = 50$ healthy males with $n_y = 52$ prostate cancer patients. Figure 1 shows a histogram and a QQ-plot of a subset of these p -values that are assumed to come from H_0 . In particular, Figure 1 reports p -values that are greater than $1/m$. This threshold is the largest Bonferroni bound that is also one of the most conservative bound in MHT (see details in Appendix A.2). In this case, p -values are obtained from the two sample t -tests (unequal variance) applied to each gene individually. As we can see from the histogram and the corresponding QQ-plot, p -values are not distributed as a $U(0, 1)$. Therefore, for these data, the uniformity assumption is questionable and cannot be further assumed.

When such a sampling null distribution of the p -values, denoted by f_0 , cannot be assumed to be $U(0, 1)$, then $\underline{B}(p)$ could be improved with another BF lower bound $\underline{B}^*(p)$. Providing $\underline{B}^*(p)$ is the aim of this work. Essentially, the idea is to substitute the theoretical null distribution $U(0, 1)$ with an empirical null distribution in the same spirit of [2, 5]. The estimation of f_0 , namely, $\hat{f}_0$, can be obtained by assuming to have observed enough p -values under the null hypothesis to precisely estimate f_0 . This is possible, for instance, in MHT where m hypotheses are tested and very few p -values, say m_1 , come from H_1 . That is

Figure 1. Prostate cancer data. p -values from the two sample t -tests: (a) Histogram of a subset of them ($p > 1/m$). (b) QQ-plot of p -values compared with the theoretical $U(0, 1)$.

$m_1 \ll m_0$, where m_0 , in some applications as RNA-seq analysis or microarray experiments, the number of p -values under H_0 , $m_0 = m - m_1$, is of the order of hundred of thousands, enough to obtain a reliable estimation of f_0 .

Organization of this paper is straightforward: Section 2 provides the proposed new calibration, $\underline{B}^*(p)$, of p -values taking advantage of the estimation of an empirical null distribution in MHT; the new calibration is compared with $\underline{B}(p)$. This comparison provides a measure of reliability of the original $\underline{B}(p)$ with respect to non-uniform distribution f_0 . This will be explored in general and for the Behrens–Fisher problem under model misspecification which is a common one in MHT with n finite. Section 3 illustrates several applications of the new calibration in real problems related to the data analysis of microarray and of RNA-seq experiments. Conclusions are left to Section 4. In order to keep the treatment as intuitive as possible, especially for applied scientists, many technical matters are left to the Appendix.

2. The calibration

Instead of assuming that the p -value null distribution f_0 is the $U(0, 1)$, we assume that this is a Beta distribution, depending on an unknown parameter $\xi_0 > 0$, so the density under H_0 is $f_0(p|\xi_0) = \text{Beta}(\xi_0, 1) = \xi_0 p^{\xi_0-1}$, where $\xi_0 > 0$. This distribution allows us to maintain simple formulas while accommodating departures from the theoretical null distribution $U(0, 1)$, obtained for $\xi_0 = 1$. In MHT, we suppose that a set of m_0 (a subset of the total m) p -values, $p_1, \dots, p_i, \dots, p_{m_0}$ are observed under H_0 , so it is possible to calculate the maximum likelihood estimation of ξ_0 , $\hat{\xi}_0 = -m_0/S$, where $S = \sum_{i=1}^{m_0} \log p_i$. It is important to note that in MHT with m tests, we can consider $m_0 = m$ or we can use a trimmed estimator, by considering only $p_i > \tilde{p}$, where $\tilde{p}$ can be some suitable threshold that refers to all p -values that will never lead to H_0 rejection for testing jointly m hypotheses. The continuity of $f_0(p|\xi_0)$ for $p < \tilde{p}$ will approximate the null for these small p -values. A good indication to fix $\tilde{p}$ may arise by considering that, under a Uniform calibration and a Bonferroni correction, p -values under the alternative are expected to be smaller than $\tilde{p} = 1/m$. By applying the restriction $p < \tilde{p}$ only for estimating ξ_0 , can it be objected that $f_0(p|\xi_0)$ has support on $p \in [\tilde{p}, 1]$ instead of on $p \in [0, 1]$. However, this would involve a slightly different normalization constant of $f_0(p|\xi_0)$ which will complicate the proposed formulas and will be practically negligible for m large.

Given the estimated empirical null distribution $f_0(p|\hat{\xi}_0)$, and using the same hypothesis testing as in the SBB calibration, see details in Appendix A.2, the new calibration for the lower bound of the BF is

$$\underline{B}^*(p) = \begin{cases} -\hat{\xi}_0 p^{\hat{\xi}_0} e \log p & p < e^{-1/\hat{\xi}_0}, \\ 1 & \text{otherwise} \end{cases}. \quad (3)$$

This calibration can also be expressed in terms of the lower bound of the null hypothesis posterior probability, $\underline{\alpha}^*(p)$, assuming 1/2 prior probability for each hypothesis, $\pi(H_0) = \pi(H_1) = 1/2$, or in terms of the lower bound of the frequentist type I error, that is,

$$\underline{\alpha}^*(p) = 1/(1 + \underline{B}^*(p)). \quad (4)$$

Note that for $S = -m$, $\hat{\xi}_0 = 1$ and $\underline{B}^*(p) = \underline{B}(p)$, we expect this only when T is ancillary or $G_0(t)$ is known exactly. What matters here is that $\underline{B}^*(p)$ is a generalization of the SBB calibration, justified under the assumption that observed p -values under the null may indicate that its sampling null distribution could not be the theoretical uniform distribution.

In order to further investigate the comparison of the proposed calibration against the SBB one, it is interesting to compare the empirical null distribution of the p -value with respect to the theoretical uniform. While this involves some technical details related to the reparametrization, we skip them in the Appendix and summarize the main features of the proposed calibration as follows:

- i) When p -values are not uniformly distributed under the null, their interpretation in terms of the least probability of the null (the conditional type I error) changes significantly according to the true null sampling distribution.
- ii) The proposed calibration allows us to increase (decrease) the minimum probability of the null hypothesis, $\underline{\alpha}^*(p)$, when many small (large) p -values are expected to be observed under the null.

Combining these two features, it becomes clear that without knowing the sampling null distribution, it is not possible to calculate the minimum probability of the null hypothesis given the data. This becomes important when the uniform distribution cannot be assumed as shown, for instance, in the following examples. Finally, to manage problems associated with MHT, it is still necessary to consider the appropriate techniques. In particular, we apply non-adaptive MHT procedure to $\underline{\alpha}(p)$ and $\underline{\alpha}^*(p)$ and for sake of comparison, the Bonferroni and the Benjamini–Hochberg (BH) [7] are considered. By no means do we state that such MHT procedures represent the state of the art of MHT, although they certainly are popular choices in MHT. A broad review of the most popular MHT literature is beyond the scope of this work. Bonferroni and BH procedures directly compare the type I error of rejecting the null with respect to a threshold that accounts for the multiplicity of the tests in order to bound false discovery rate (FDR) and/or family wise error rate (FWER). So it is worthwhile comparing such thresholds against $\underline{\alpha}(p)$ and $\underline{\alpha}^*(p)$. In particular, we consider the nominal level for FWER and FDR equal to 0.05, namely, we report as discoveries those tests whose ordered $\underline{\alpha}^*(p)$ and/or $\underline{\alpha}(p)$ are less than $0.05/m$ for the Bonferroni and less than $0.05i/m$ for the BH procedure where i represents the order position along the m tests, being $i = 1$ the smallest value and $i = m$ the largest one.

2.1. Example: the Behrens–Fisher problem for normal populations

Consider the usual Behrens–Fisher problem in which the mean of two independent normal populations are compared assuming that all parameters are unknown. More formally, let m be the number of tests and denote with $x_{m \times n_x}$ the outcome in population X with n_x replications and $y_{m \times n_y}$ the outcome in population Y with n_y replications. Let $X_i \sim N(\mu_{X_i}, \sigma_{X_i}^2)$ and $Y_i \sim N(\mu_{Y_i}, \sigma_{Y_i}^2)$ for $i = 1, 2, \dots, m$. The set of m hypotheses testing, for $\sigma_{X_i}^2 > 0$, $\sigma_{Y_i}^2 > 0$ unknown, is the following:

$$\{H_{0i} : \mu_{X_i} = \mu_{Y_i} = \mu_i \text{ versus } H_{1i} : \mu_{X_i} \neq \mu_{Y_i}, \forall \sigma_{X_i}^2 > 0, \forall \sigma_{Y_i}^2 > 0\}, i = 1, \dots, m.$$

Suppose that p -values come from the usual *Student t*-test for two samples with the Welch correction for $\sigma_{X_i}^2 \neq \sigma_{Y_i}^2$. This is just an asymptotic correction that does not guarantee uniform p -values in finite samples.

In a small simulation study, we compare SBB calibration against the one proposed by looking at the differences between $\underline{\alpha}(p)$ and $\underline{\alpha}^*(p)$ calculated over p -values for the aforementioned hypotheses testing

Figure 2. Behrens–Fisher problem: outcome from one simulated experiment with $n = 5$. The grey points are p -values under the null, while the black points are under the alternative. (a and b) are QQ-plots of observed p -values against the $U(0, 1)$ and the empirical null which is $Beta(\hat{\xi}_0 = 1.13, 1)$, respectively. (c and d) report the comparison of $\underline{\alpha}(p)$ and $\underline{\alpha}^*(p)$ in terms of values and in terms of differences separated for p -values under the null and the alternative.

H_{0i} against H_{1i} , for $i = 1, \dots, m$. Suppose $m = 10,000$ tests with sample sizes $n_x = n_y = 5$. We assume that 1% of the tests (100 tests) comes from the alternative, with $\mu_{X_i} = 2$ and $\mu_{Y_i} = 0$, while for the rest of hypotheses, we consider $\mu_{X_i} = \mu_{Y_i} = 0$ and homoscedasticity $\sigma_{X_i}^2 = \sigma_{Y_i}^2 = 1$, which is the most favorable situation for the Welch correction. Results for the simulated data are reported in Figure 2. First of all, we can see that because of the small sample, the Welch correction provides non-uniform p -values as shown in Figure 2 (a). In particular, the Welch correction in this example produces conservative p -values and the estimated empirical null is the $Beta(\hat{\xi}_0 = 1.13, 1)$ which fits the observed p better, as reported in Figure 2 (b). Therefore, there exist differences between the SBB calibration and the one proposed as illustrated in Figure 2 (c). In these series of simulated MHT, bounds α^* are lower than those with SBB calibration as expected for conservative distributions (see Appendix A.2, the last row of Table AI). However, we seek an improvement in separating the evidence from null versus alternative hypotheses, and this is reported in the vertical axis of Figure 2 (d). We can see larger differences under the alternative than under the null. This increases the separation, in terms of calibrated evidence among p -values coming from H_1 and H_0 and thus the chance of detecting the two sets of tests.

Such a result, although referring to simulated data, basically illustrates the difference in assessing the evidence of those tests from the alternative (the black points in Figure 2) with respect to those under the null (the grey points). In fact, differences in recalibrating p -values under the alternative are much less than those under the null, and this may affect an MHT procedure that can be based either on $\underline{\alpha}(p)$ or the newly proposed $\underline{\alpha}^*(p)$, as reported in the following applications to real data.

In order to emphasize this separation effect and also the usefulness in assessing MHT based on $\underline{\alpha}(p)$ and more importantly on $\underline{\alpha}^*(p)$, we reported results of an extensive simulation study in which the Student t -test with the Welch correction has been applied to two independent gamma populations. The considered set up is as follows: $m = 100,000$, $m_1 = 100$, $n = 10, 100$, $Y_i, X_i \sim \text{Gamma}(1, 1)$ under H_0 and $Y_i \sim \text{Gamma}(5, 1)$ under H_1 ; that is, the mean and the variance are larger by a factor of 5 under the alternative. Summarizing, in this setup, different analysis features are considered:

1. Model misspecification: normal populations are be considered, while observations do not come from a normal population. This is a relevant misspecification feature in microarray analysis with

Figure 3. Behrens–Fisher problem: outcome of FDR and false non-rejection rate (log 10 scale) in 1 million simulated multiple testing experiments under the following set up: $m = 100,000$, $m_1 = 100$, $n = 10, 100$, $Y_i, X_i \sim \text{Gamma}(1, 1)$ under H_0 and $Y_i \sim \text{Gamma}(5, 1)$ under H_1 . The point represents the mean quantity in log 10 scale in 1 million simulations, while vertical red bars represent ± 3 standard error of the mean. [Colour figure can be viewed at wileyonlinelibrary.com]

small sample sizes. Such a feature leads to p -values that are not distributed $U(0, 1)$, and this affects the usual MHT procedures in that the control of FDR is no longer guaranteed as already illustrated in [8];

2. The well-known BF consistency allows to $\underline{\alpha}(p), \alpha^*(p) \rightarrow 0$ for $n \rightarrow \infty$ faster than the original p -value, p , with larger separation achieved with $\alpha^*(p)$ rather than with $\underline{\alpha}(p)$;
3. High precision in estimating $\hat{\xi}_0$ for m large.

The simulation setup has been replicated one million times to estimate the error in approximating the actual FWER and FDR, with the Bonferroni and the BH [7] procedures, respectively. These have been applied to the original p -values, p , as well as to $\underline{\alpha}(p)$ and $\alpha^*(p)$ at the nominal level of 10%. Both procedures control the FDR and the Bonferroni procedure, which is more conservative, also controls the FWER. No control over the false non-rejection rate is guaranteed.

Figure 3 still shows that differences exist in calibrating null and alternative hypotheses. First of all, it is clear that using the conditional type I error in such MHT procedures produces less FDR with respect to the original p -value. This is a result that applies along all the setup either when controlling the FWER or the FDR. In small samples, differences between $\underline{\alpha}(p)$ and $\alpha^*(p)$ are less clear, while in large samples, the BF consistency argument makes the FDR less with $\alpha^*(p)$ than with $\underline{\alpha}(p)$ supporting the need for a proper p -value calibration.

To finally allow applied analysts to investigate their own data in a very simple and intuitive environment, a web application, implementing the proposed method, is available at https://stefano-cabras.shinyapps.io/p-value_calibration/ (see also supplementary materials).

3. Application

In this section, we compare $\underline{\alpha}^*(p)$ and $\underline{\alpha}(p)$ in the analysis of a series of microarray and RNA-seq experiments.

Figure 4. Results from the analysis of prostate cancer data with $\tilde{p} = 1/m$. (a) QQ-plot of original p -values compared with the theoretical $U(0, 1)$. (b) The density of the empirical null $\hat{f}_0$. (c) QQ-plot of empirical distribution of p -values contrasted with $\hat{f}_0$. (d) Box-plots of $\alpha^*(p)$ and $\alpha(p)$.

Figure 5. Results RNA-seq experiment. QQ-plot of p -values supposed to come from the null model according to [9].

The first is a large microarray study also analyzed, with different approaches, in [6] and [2], and evidence comes from p -values of the usual *Student t*-tests with the Welch correction as illustrated earlier.

The second example is a post-analysis of a more recent outcome of an RNA-seq experiment.

Finally, $\alpha^*(p)$ is estimated in both experiments for those $p_i > \tilde{p} = 1/m$, which is a quite conservative threshold in MHT. A sensitivity analysis with respect to this threshold is provided in the supplementary material. Essentially, for m , large results are insensitive to $\tilde{p}$.

3.1. Microarray: prostate data

In this study of gene expression levels for prostate cancer data, [6], $m = 6033$ genes with $n_x = 50$ healthy males are compared with $n_y = 52$ prostate cancer cases. Results are shown in Figure 4.

From Figure 4 (a), we can observe that the p -value null distribution is not very different from the theoretical $U(0, 1)$ suggesting that there are possibly very few genes that come from the alternative. In fact, the estimated empirical null $f_0(p|\hat{\xi}_0 = 0.85)$ does not differ too much from the $U(0, 1)$, being $\hat{\xi}_0 = 0.85$ near the theoretical value of 1. Of course, given that m is large, such an estimation differs significantly from 1. That is, according to the observed p -values, the theoretical sampling null distribution, $U(0, 1)$, is not compatible with the data. This is the main argument that motivates analysis of this data set with the empirical null approach illustrated in [2] but applied to the T scale rather than to the p -value scale as instead proposed here. The empirical null density in Figure 4 (b) suggests that there are too many small p -values with respect to that expected under $U(0, 1)$, and thus, the new empirical null fits almost all p -values better as reported in Figure 4 (c). This finally produces values of $\alpha^*(p)$ that are much larger than the corresponding $\alpha(p)$ suggesting that there should be much fewer differentially expressed genes than those that can be obtained using the conditional type I error with $\alpha(p)$. In fact, according to the Bonferroni and BH procedures, with $\alpha(p)$, one gene should be considered as differentially expressed, while if we account for the fact that the p -value null distribution is not uniform, then no gene should be pointed to as differentially expressed.

3.2. An RNA-seq experiment: Bovine macrophage response to *Mycobacterium bovis* infection

In this study, the raw data consist of 3.6 trillion reads of RNA sequences in order to generate counts of identical sequences that are supposed to represent the abundance of target sequences in the mRNA parts under study. Counts are collected for the two biological populations under comparison: bovines infected and non-infected by *Mycobacterium bovis* [9]. Such counts are then used to evaluate the differential gene expressions between such biological populations (see for instance, [10]). After data normalization, there are $m = 11131$ genes that have been compared for differential expressions and their corresponding p -values have been collected.

In [9] 2584 genes have been declared as differentially expressed, namely those with adjusted p -value with the BH procedure less than 0.05. However, we can see from Figure 5 that such p -values are not uniformly distributed as expected. Therefore, in order to properly evaluate their evidence, especially in MHT, they must be calibrated with respect to another non-uniform distribution under the null.

Results of the proposed calibration are reported in Figure 6 .

From Table I we have that 2584 genes is too large a set of differentially expressed genes and it is in fact much smaller if the conditional type I error $\alpha(p)$ is considered and even less if considering $\alpha^*(p)$.

In fact, given that the theoretical null distribution of the p -value is not $U(0, 1)$, as we can see from Figure 6 (a), we should consider the empirical null instead, that is the Beta density in Figure 6 (b) and (c) with $\tilde{p} = 1/m$. This finally ends up in fewer genes whose conditional type I error is smaller, that is $\alpha^*(p)$ is larger than its corresponding $\alpha(p)$ as shown in Figure 6 (d) and also in Table I in the last three rows. This is also consistent with analysis of this data performed in [11] under a full Bayesian approach on modeling the observed p -values.

Note that in this experiment the choice of $\tilde{p}$ is important. It is possible to observe in Figure S2 that choices of $\tilde{p} \in \{0, 1/m, 2/m\}$ produce different results in terms of different empirical null densities and number of discoveries (Table I). In fact, among the estimated Beta densities the one estimated with $\tilde{p} = 1/m$ fits the p -values just right (Figure S2, left column). This turns into not very different values of $\alpha^*(p)$ (Figure S2, right column), but enough to change the number of discoveries (see the last three rows of Table I) especially from $\tilde{p} = 0$ to $\tilde{p} = 1/m$.

Table I. RNA-seq experiment.

		Bonferroni (FWER)	BH (FDR)
$\alpha(p)$		728	1490
$\alpha^*(p)$	$(\tilde{p} = 0)$	72	154
	$(\tilde{p} = 1/m)$	339	675
	$(\tilde{p} = 2/m)$	355	703

Number of non-null RNA sequences according to the Benjamini–Hochberg (BH) procedure and Bonferroni procedure, both applied to $\alpha(p)$ and $\alpha^*(p)$, respectively. Values are separated for different choices of $\tilde{p}$ where $m = 11131$.

Figure 6. Results RNA-seq experiment with $\tilde{p} = 1/m$. (a) QQ-plot original of p -values against the theoretical $U(0, 1)$. (b) The density of the empirical null $\hat{f}_0$. (c) QQ-plot of empirical distribution of p -values against $\hat{f}_0$. (d) Box-plots of $\underline{\alpha}^*(p)$ and $\underline{\alpha}(p)$.

4. Conclusions

This paper proposes a generalization of the SBB calibration of the p -values with respect to BF lower bound and conditional type I error rates $\underline{\alpha}(p)$. Such a generalization simply consists in relaxing the $U(0, 1)$ calibration for the p -values under the null and letting the observed p -values be used to estimate an empirical null distribution in the same spirit of [2]. This is possible in MHT problems in which many p -values are generally observed and very few come from alternative hypotheses. We are not aware of other inferential settings, but we think that also in meta-analysis, the proposed approach could be realistic. With many tests supposed to be observed under the null, it is thus possible to estimate the empirical null model reliably with many observations for estimating the parameter ξ_0 , which indexes the assumed class of p -value sampling null distributions. Such a class may be viewed as a restrictive class, but it is considered in order to keep the level of calculus simple, in the same spirit of [3], where the restriction of the empirical null density to the family of $\text{Beta}(\xi_0, 1)$ densities provides a simple formula. This choice is made with the hope of extending the calibration approach proposed here of the p -value with respect to the conditional evidence of the null, an approach also advocated in [1]. However, the popularity of new calibrations can still be improved, without the necessity of simple formulas by implementing an R package that deals with a broader class of empirical null densities as in [2]. This would be an interesting line of research that is beyond the scope of the present work. Finally, we always recommend that the process of questioning or rejecting null hypotheses be performed using the appropriate MHT techniques and also with a very careful sensitivity analysis on the data.

Appendix A

A.1. Technical details Section 1

To obtain the calibration in (1) and (2) in [3] and [4], the following model selection (or hypothesis testing) problem is considered:

$$\begin{cases} H_0 : & p \sim f_0 \\ H_1 : & p \sim f_1 \end{cases}, \quad (\text{A.1})$$

where f_0 and f_1 represent sampling distributions of the p -value under the null and alternative hypotheses, respectively.

Considering the distribution of the test statistic, T known, $G_0(t)$, then $f_0 \equiv U(0, 1)$ and a p -value is said to be *calibrated* with respect to the Uniform distribution.

The alternative densities f_1 in [3] and [4] are assumed to be in the class of all $\text{Beta}(\xi, 1)$ densities for $0 < \xi \leq 1$, that is all p -values densities that concentrate at 0 under H_1 . We also restrict analysis to this class of density in the proposed calibration because it seems to us realistic and also allows for a simple formula. Under such assumptions in [3] and [4], (A.1) becomes

$$\begin{cases} H_0 : & p \sim U(0, 1) \\ H_1 : & p \sim f(p|\xi) = \xi p^{\xi-1} \end{cases}, \quad (\text{A.2})$$

The Bayes factor (BF) of H_0 against H_1 for any given continuous prior density $\pi(\xi)$ for the alternative H_1 is

$$B_\pi(p) = \frac{\mathbb{I}_{[0,1]}(p)}{\int_0^1 f(p|\xi) \pi(\xi) d\xi},$$

where $\mathbb{I}_A(x)$ is the usual indicator function of $x \in A$. This leads to the following BF lower bound, for every $\pi(\xi)$,

$$\underline{B}(p) = \inf_{\forall \pi} B_\pi(p) = \frac{1}{\sup_{\xi} \xi p^{\xi-1}} = \begin{cases} -ep \log p & p < e^{-1} \\ 1 & \text{otherwise} \end{cases}.$$

The lower bound of the null hypothesis posterior probability, $\underline{\alpha}(p)$, assuming 1/2 prior probability for each hypothesis, is

$$\underline{\alpha}(p) = 1/(1 + \underline{B}(p)).$$

$\underline{\alpha}(p)$ is also the lower bound of the frequentist type I error probability ([3]).

A.2. Technical details Section 2

Given the estimated empirical null distribution $f_0(p|\hat{\xi}_0)$, and using the same hypothesis testing as in the SBB calibration, (A.2) becomes

$$\begin{cases} H_0 : & p \sim f_0(p|\hat{\xi}_0) \\ H_1 : & p \sim f(p|\xi) = \xi p^{\xi-1} \end{cases}, \quad (\text{A.3})$$

with BF of H_0 against H_1 equal to

$$B_\pi^*(p) = \frac{\hat{\xi}_0 p^{\hat{\xi}_0-1} \mathbb{I}_{[0,1]}(p)}{\int_0^1 f(p|\xi) \pi(\xi) d\xi} \quad (\text{A.4})$$

which leads to the following new BF lower bound for every prior $\pi(\xi)$

$$\underline{B}^*(p) = \inf_{\forall \pi} B_\pi^*(p) = \frac{\hat{\xi}_0 p^{\hat{\xi}_0-1}}{\sup_{\xi} \xi p^{\xi-1}} = \begin{cases} -\hat{\xi}_0 p^{\hat{\xi}_0} e \log p & p < e^{-1/\hat{\xi}_0} \\ 1 & \text{otherwise} \end{cases}.$$

Assuming equal probability for each hypothesis, $\pi(H_0) = \pi(H_1) = 1/2$, the lower bound of the null hypothesis posterior probability, $\underline{\alpha}^*(p)$, which agrees with the lower bound of the frequentist type I error is

$$\underline{\alpha}^*(p) = 1/(1 + \underline{B}^*(p)).$$

A way to compare the SBB calibration and the one proposed here is through a reparametrization $D = \hat{\xi}_0 - 1$, which represents a signed distance whose values are interpreted as illustrated in Table AI.

Table A1. Interpretation of the signed distance D , where the term conservative (anti-conservative) refers to the null distribution of the p -value when it concentrates in high (low) values.

D	$\hat{\xi}_0$	Interpretation of the test
$(-1, 0)$	$(0, 1)$	Anticonservative p -value
0	1	$U(0, 1)$ calibrated p -value
$(0, \infty)$	$(1, \infty)$	Conservative p -value

Figure A1. Graph of $\underline{\alpha}^*(p, D)$ for different p -values: 0.001, 0.01 and 0.05. Horizontal lines represent the SBB $\underline{\alpha}(p)$ calibration of the corresponding p -values.

For $D > 0$ ($D < 0$), the p -value is conservative (anticonservative) as its sampling null distribution tends to accept (reject) null hypotheses more than expected under the uniform calibration ($D = 0$).

The effect of discrepancy between the empirical null and the uniform calibration can be expressed in terms of $\underline{\alpha}^*(p, D)$ as a function of D for some fixed p -value. This amounts to reparametrizing (3) with $\hat{\xi}_0 = D + 1$. For different p -values, the corresponding graphs of $\underline{\alpha}^*(p, D)$ are reported in Figure A1 along with the corresponding SBB calibration in terms of $\underline{\alpha}(p)$.

At least two features of the proposed calibration can be retrieved from Figure A1:

- (i) Differences of the empirical null with respect to the theoretical uniform distribution matter in calibrating the p -values because the SBB calibration does not account for departure from the theoretical $U(0, 1)$.
- (ii) For $D > 0$ ($D < 0$), the new lower bound of the probability of the null is smaller (greater) than that of SBB, and this is because the null distribution of the p -value is conservative (anticonservative) and low p -values should be interpreted as greater (smaller) evidence against the null in relation to their empirical null distribution.

The maximum of $\underline{\alpha}^*(p, D)$ represents those p -values for which the posterior probability on H_0 equals its prior $\pi(H_0)$. As this maximum is achieved at $p = e^{-1/\hat{\xi}_0}$, it is immediate to check that with just one observed p -value, we have the identity $p = p$ and $\underline{\alpha}^*(p) = \pi(H_0)$. Practically, whatever the observed p is and without knowing the p -value sampling null distribution, in this case, it is not possible to update the prior probability of the null hypothesis.

Acknowledgements

Stefano Cabras and María Eugenia Castellanos have been partially funded by Ministerio de Economía y Competitividad of Spain under grants MTM2016-77501-P, ECO2015-66593-P, RYC-2012-11455.

References

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 2016; **70**(2):129–133.
2. Efron B. Microarrays, empirical Bayes and the two-groups model. *Statistical Science* 2008; **23**(1):1–22.
3. Sellke T, Bayarri MJ, Berger JO. Calibration of p-values for testing precise null hypotheses. *American Statistician* 2001; **55**(1):62–71.
4. Bayarri MJ, Berger JO. P values for composite null models. *Journal of the American Statistical Association* 2000; **95**(452):1127–1142.
5. Efron B. Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association* 2004; **99**:96–104.
6. Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D'Amico A, Richie J, Lander E, Loda M, Kantoff P, Golub T, Sellers R. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 2002; **1**: 302–309.
7. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995; **57**:289–300.
8. Cabras S. A note on multiple testing for composite null hypotheses. *Journal of Statistical Planning and Inference* 2010; **140**:659–666.
9. Nalpas NC, Park SDE, Magee DA, Taraktsoglou M, Browne JA, Conlon KM., Rue-Albrecht K, Killick KE, Hokamp K, Lohan AJ, Loftus BJ, Gormley E, Gordon SV, MacHugh DE. Whole-transcriptome, high-throughput RNA sequence analysis of the bovine macrophage response to *Mycobacterium bovis* infection *in vitro*. *BMC Genomics* 2013; **14**:230. <https://doi.org/10.1186/1471-2164-14-230>.
10. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* 2013; **14**(9):R95.
11. Cabras S. A Markov chain representation of the multiple testing problem. *Statistical Methods in Medical Research* 2016 DOI:10.1177/0962280216628903.

Supporting information

Additional supporting information may be found online in the supporting information tab for this article.