Gomez-Villegas, M. A. & González-Pérez, B. (2018) Multiple Hypothesis tests: A Bayesian approach. in E. Gil. E. Gil et al. The Mathematics of the Uncertain. Switzerland: Springer.

Multiple Hypothesis Tests: A Bayesian Approach

Miguel A. Gómez-Villegas and Beatriz González-Pérez

Abstract Multiple hypothesis tests is a topic which has recently shown a major expansion, mainly due to the expansion of the methodology developed in connection with genomics. These new methods allow scientists to handle simultaneously thousands of null hypotheses. The frequentist approach to this problem consists of using different error measures in testing so that to ensure the Type I error remains below a desired level. This paper introduces a parametric Bayesian analysis to determine the hypotheses to be considered as being significant (i.e., useful) for a posterior deeper analysis. The results are to be compared with the frequentist methodology of the false discovery rate (FDR). Differences between both approaches are shown by means of simulation examples.

1 Introduction

This article is my memory to Pedro Gil Álvarez who was my professor circa 1970 at the Faculty of CC. Mathematics at the Complutense University.

I always thought Pedro was a very intelligent person. When I was in my 4th and 5th year of College he was in charge of the labs of the most diverse subjects in the field of statistics. Later, I realized that Teaching Assistants were scarce in the department. As a consequence, the professors who were at the beginning of their careers had to perform a remarkable effort to deal with students who were just at the initial level of statistics. This reinforced my feeling that he was very smart.

In the field of simultaneous inference, multiple hypothesis testing deals with the testing of more than one hypothesis at time. A single hypothesis test can be described as follows

M. A. Gómez-Villegas (✉) · B. González-Pérez
Departamento de Estadística e Investigación Operativa I. Facultad de
Ciencias Matemáticas, Instituto de Matemática Interdisciplinar (IMI),
Universidad Complutense de Madrid, 28040 Madrid, Spain
e-mail: ma.gv@mat.ucm.es

B. González-Pérez
e-mail: beatrizg@mat.ucm.es

E. Gil et al. (eds.), *The Mathematics of the Uncertain*, Studies in Systems, Decision and Control 142, https://doi.org/10.1007/978-3-319-73848-2_19

$$H = 0 : \theta \in \Theta_0 \quad \text{versus} \quad H = 1 : \theta \in \Theta_1 \tag{1}$$

with $\Theta_0 \cap \Theta_1 = \emptyset$. A statistic, $T(X)$, is observed and a value $T(x) = t$ is obtained.

From the frequentist point of view, the null hypothesis will be rejected if the observed value, t, is over a certain threshold. This threshold, which is arbitrary, settles a certain rejection region, Γ, in such a way that if $t \in \Gamma$ then $H = 0$ is rejected while if $t \notin \Gamma$, then $H = 0$ is accepted. The rejection region defines the Type I error, that is to reject the null hypothesis when it is true, $\theta \in \Theta_0$ but $t \in \Gamma$.

When testing a single hypothesis as (1), an acceptable maximum Type I error probability is specified and the conclusions are obtained based on a statistic which meets this specification. Then, the maximum Type I error probability is fixed at a certain level, which is known as significance level, α

$$\sup_{\theta \in \Theta_0} Pr(T \in \Gamma_\alpha | \theta) = Pr(T \in \Gamma_\alpha | H = 0) = \alpha \tag{2}$$

and a frequentist measure of the evidence against the null hypothesis is the p–value, defined as the minimum false positive rate at which an observed statistic can be called significant,

$$p - value(t) = \sup_{\theta | H = 0} Pr(T \in \Gamma_t | H = 0) \tag{3}$$

where Γ_t is the critical region for $T = t$. Alternatively, the probability that, the statistic is as or more extreme than the observed one, t, under the null hypothesis,

$$p - value(t) = Pr(|T(X)| \geq |t| | H = 0)$$

can be used as test statistics (see Lehmann and Romano [8, p. 63]).

But when many hypothesis are tested, to fix an individual Type I error probability for each one may have consequences if the set of hypothesis are evaluated as a whole. A review about multiple hypothesis testing can be seen in Shaffer [10].

The question is, basically, whether the probability of a false positive increases with the number of tests. For example, if the significance level is fixed at 0.05 for each test and a set of 100 tests are evaluated, the expected number of false positives is 5. Then, 5 hypothesis will be rejected simply by chance. The level 0.05 has been widely used in the literature since Fisher proposed it, and its intense use has produced basically correct scientific inferences. Otherwise it would not have remained as a reference level so long. But the 0.05 level was applied to a single hypothesis, not to a great number of simultaneous hypothesis; this is the reason for introducing measures of evidence that take into account all the hypothesis which are tested simultaneously.

Multiple hypothesis testing has been widely used in the past in different fields as Shaffer [10] pointed out. Recently, the field of genomics, and in particular the DNA microarray experiments where thousands of hypothesis can be tested simultaneously, have influenced the revitalization of the procedures of the multiple hypothesis tests, see Dudoit et al. [4] In this context, consider m hypothesis tests

$$H_i = 0 \quad \text{versus} \quad H_i = 1, \quad i = 1, \ldots, m \tag{4}$$

and we want to test the m null hypothesis simultaneously. Benjamini and Hochberg [2] propose Table 1 to summarize the problem.

Table 1 Multiple hypothesis

	Number accepted	Number rejected	
$H = 0$	U	V	m_0
$H = 1$	T	S	m_1
	W	R	m

In Sect. 2 we have summarized some of the frequentist procedures to test m null hypothesis simultaneously, as in (4). In Sect. 3 we have proposed two Bayesian approaches for testing (4) and we compare, for simulated data, the results obtained for these Bayesian methods with the frequentist results. Section 3 includes a hierarchical model. Finally, Sect. 4 contains some conclusions and comments.

2 Frequentist Procedures

2.1 Type I Error Rates

Shaffer [10] picked out the generalizations of the Type I error described above to the multiple testing problems: the family wise error rate, FWER, the per–comparison error rate PCER, the per-family error rate, PFER, and the false discovery rate, FDR. This last one was introduced by Benjamini and Hochberg [2].

In any case, the procedure to carry out a multiple hypothesis test consists of controlling a particular Type I error rate at a certain level α and producing a list of R rejected hypothesis. If the level α is fixed to control the Type I error rate only when all the null hypothesis are true, $m_0 = m$, one speaks of weak control, whereas strong control referrers to the control of the Type I error rate under any posible combination of the true and false null hypothesis. See Shaffer [10] and Dudoit et al. [4] for more details about this setting.

2.2 p-Values

As defined above, the p-value $p_i(t_i)$ for a single hypothesis H_i can be viewed as the level of the test at which the hypothesis H_i would be rejected, given the value of a statistic $T_i = t_i$. The smaller the p-value, $p_i(t_i)$, the stronger the evidence against the null H_i. With a fixed significance level, α, rejecting H_i when $p_i \leq \alpha$ assumes that the Type I error rate is controlled at level α.

The concept of the p-value can be extended to the multiple testing problem under the concept of the adjusted p-value, see Dudoit et al. [4]. Given a test procedure, for example a FDR procedure (see Benjamini and Hochberg [2]) defined as in (6), the adjusted p-value for a single hypothesis H_i is defined as

$$\widetilde{p}_i = \inf\{\alpha \in [0, 1] : H_i \text{ is rejected at nominal FDR} = \alpha\} \tag{5}$$

in words, the nominal level of the entire test procedure at which the hypothesis H_i would be rejected, given the values of all the test statistics.

Westfall and Young [13] estimated the adjusted p-values by resampling methods. A recent discussion about p-values can be seen in Wasserstein and Lazar [12].

2.3 *The False Discovery Rate*

Benjamini and Hochberg [2] introduced this concept, less conservative than the others, to control the expected proportion of Type I errors among the rejected hypothesis. So, the FDR is defined as, from Table 1,

$$FDR = E\left[\frac{V}{R}\right] Pr(R > 0). \tag{6}$$

If $R = 0$, then $FDR = 0$. Benjamini and Hochberg [2] derived a procedure for strong control of the FDR for independent test statistics. This procedure for control of the FDR at level α can be resumed as follows:

- The observed p-values are computed and ordered: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
- Compute $R_{BH} = \max\{i : p_{(i)} \leq \alpha(i/m)\}$.
- Reject null hypothesis corresponding to $p_{(1)}, \ldots, p_{(R_{BH})}$. If R_{BH} does not exist, no hypothesis is rejected.

Benjamini and Yekutieli [3] showed that the procedure above controls the FDR at level α, under certain conditions of dependency.

The adjusted p-values corresponding to this control are

$$\widetilde{p}_{(i)} = \min_{j=i,\ldots,m}\left\{\min\left(\frac{m}{j}p_{(j)}, 1\right)\right\}. \tag{7}$$

In a microarray setting, Dudoit et al. [4] proposed the FDR controlling procedures as alternatives to other approaches. They argue that in this context one may be willing to bear a few false positives as long as their number is small in comparison to the number of rejected hypothesis.

3 A Bayesian Approach

Consider the problem of multiple hypothesis testing given in (4). As in Table 1, we denote by m_1 the unknown number of hypotheses where $H_i = 1, i = 1, \ldots, m$. In this section we propose two different Bayesian methods for testing (4).

Bayesian inference on multiple hypothesis has been widely studied. We are studying it in the same way as the approach proposed by Waller and Duncan [11], Robert [7], Barbieri and Berger [1] and Scott and Berger [9].

Our objective is to give a Bayesian estimator for m_1, for instance, the mean or the mode of the posterior density of m_1. We denote θ_i by the prior probability of the null $H_i = 0, \theta_i = P(H_i = 0)$, and consequently $1 - \theta_i = P(H_i = 1), i = 1, \ldots, m$, and supposing that the m hypothesis are independent, then $H_i|\theta_i \sim Bernoulli(1 - \theta_i)$ and $m_1 = \sum_{i=1}^{m} H_i$. An initial Bayesian approach is to assume that all θ_i are equal to θ. Then, $m_1|\theta \sim Binomial(m, 1-\theta)$ and we can estimate m_1 with $\hat{m}_1 = m(1-\hat{\theta})$, where $\hat{\theta}$ is an estimator of θ (a location parameter of the posterior density of θ).

For each hypothesis H_i, a vector $T_i = (X_{i1}, \ldots, X_{in})$ is observed. Suppose that for all i, under $H_i = 0$ the density is $f(t|0)$, while under $H_i = 1$ the density is $f(t|1)$. Thus, the observations T_i (assume i.i.d. random variables) come from a mixture of both densities:

$$\begin{aligned} f(t_i|\theta) &= f(t_i|H_i = 0)Pr(H_i = 0|\theta) + f(t_i|H_i = 1)Pr(H_i = 1|\theta) \\ &= \theta f(t_i|0) + (1-\theta) f(t_i|1) \end{aligned} \tag{8}$$

and the likelihood can be written as

$$f(t_1, t_2, \ldots, t_m|\theta) = \prod_{i=1}^{m} f(t_i|\theta) = \prod_{i=1}^{m} [\theta f(t_i|0) + (1-\theta) f(t_i|1)], \tag{9}$$

where $t_i = (x_{i1}, \ldots, x_{in})$.

The prior distribution for the parameter θ can be thought of as a beta distribution, $Beta(a, b)$, because of its versatility to model a density over the interval [0, 1]. Then,

$$\pi(\theta|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 \leq \theta \leq 1 \tag{10}$$

and the posterior density of θ, given $t_1, \ldots, t_m, a, b$, is given by

$$\pi(\theta|t_1, \ldots, t_m, a, b) = \frac{\pi(\theta|a, b)\prod_{i=1}^{m}[\theta f(t_i|0) + (1-\theta) f(t_i|1)]}{\displaystyle\int_0^1 \pi(\theta|a, b)\prod_{i=1}^{m}[\theta f(t_i|0) + (1-\theta) f(t_i|1)]\,d\theta} \tag{11}$$

where, in the first step, a and b are known and previously fixed.

Then, we compute and order the posterior probabilities

$$P(H_i = 0|T_i = t_i) = \frac{f(t_i|0)\theta}{f(t_i|0)\theta + f(t_i|1)(1-\theta)}, \quad i = 1, \dots, m. \tag{12}$$

These probabilities will be estimated using $\widehat{\theta}$, the estimated value of θ obtained through (11), then

$$\widehat{P}(H_i = 0|T_i = t_i) = \frac{f(t_i|0)\widehat{\theta}}{f(t_i|0)\widehat{\theta} + f(t_i|1)(1-\widehat{\theta})}, \quad i = 1, \dots, m. \tag{13}$$

Really, the Bayesian method involves computing $P(H_i = 0|T_1 = t_1, \dots, T_m = t_m)$, but the approximation proposed by (12) and (13) simplifies the simulation's task a lot.

Finally, we will use $\widehat{\theta}$ and the estimated posterior probabilities in multiple hypothesis testing using the two following methods:

Method 1: We estimate the percentage of hypothesis where $H_i = 0$ by using the mean $\widehat{\theta}$ of the posterior density (11) as an estimator of θ. In this case, $\widehat{m}_1 = m(1-\widehat{\theta})$. Then, the $\widehat{m}_1$ hypotheses with the lowest estimated posterior probabilities, see (13), will be rejected (will be declared interesting).

Method 2: We use the following Bayesian decision procedure

- Given $\widehat{\theta}$, the observed posterior probabilities are computed from (13) and ordered.
- Compute $\widehat{i} = \max\{i : \widehat{P}\left(H_{(i)} = 0|T_{(i)} = t_{(i)}\right) \leq 0.5\}$. Thresholds other than 0.5 can be used.
- Reject the null hypothesis corresponding to $t_{(1)}, \dots, t_{(\widehat{i})}$. If $\widehat{i}$ does not exist, no hypothesis is rejected.

Then, the hypotheses with the estimated posterior probabilities lower than 0.5 will be rejected.

From the Bayesian point of view, Method 2 is formally more correct than Method 1. But, experimentally, when simulations with m_1 known are run, Method 1 adjusts the results better when the hypothesis are close and the sample size, n, is small which is usually the case in these kinds of problems. Whereas if the hypothesis are not close both methods provide similar results.

The next example, used by the authors previously cited, shows how the methodology is applied to an example with a normal model.

Example 3.1 If under $H_i = 0$ the model is $N(0, 1)$ and under $H_i = 1$ is $N(1, 1)$, and n observations are taken for all $i = 1, \dots, m$, then

$$\begin{aligned} f(t_i|\theta) &= \theta \prod_{j=1}^{n} f(x_{ij}|0) + (1-\theta) \prod_{j=1}^{n} f(x_{ij}|1) \\ &= (2\pi)^{-n/2} e^{-(1/2)\sum_{j=1}^{n} x_{ij}^2} (\theta + (1-\theta)e^{n(\bar{x}_i - 1/2)}). \end{aligned}$$

Then the joint distribution of the nm observations is

$$\begin{aligned} f(t_1, \dots, t_m|\theta) &= \prod_{i=1}^{m} f(t_i|\theta) \\ &= (2\pi)^{-nm/2} e^{-(1/2)\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}^2} \prod_{i=1}^{m} (\theta + (1-\theta)e^{n(\bar{x}_i - 1/2)}). \end{aligned} \tag{14}$$

And the posterior density of θ given $t_1, \dots, t_m, a, b$, is

$$\begin{aligned} \pi(\theta|t_1, \dots, t_m, a, b) &\propto \pi(\theta|a, b) f(t_1, \dots, t_m|\theta) \\ &\propto \theta^{a-1}(1-\theta)^{b-1} \prod_{i=1}^{m} (\theta + (1-\theta)e^{n(\bar{x}_i - 1/2)}) \end{aligned}$$

whereas the posterior probability of the null is estimated by

$$\hat{P}(H_i = 0|T_i = t_i) = \left(1 + \frac{1-\widehat{\theta}}{\widehat{\theta}} e^{n(\bar{x}_i - 1/2)}\right)^{-1} \tag{15}$$

for $i = 1, \dots, m$ hypothesis, with $\widehat{\theta} = E[\theta|t_1, \dots, t_m, a, b]$.

We use Montecarlo integration to estimate $\widehat{\theta}$. For this, we simulate a random sample $\theta_1, \dots, \theta_k$ from the prior distribution $Beta(a, b)$. Then, we estimate $\widehat{\theta}$ as:

$$\begin{aligned} \widehat{\theta} = E[\theta|t_1, \dots, t_m, a, b] &= \frac{\int_0^1 \theta f(t_1, \dots, t_m|\theta)\pi(\theta|a, b)d\theta}{\int_0^1 f(t_1, \dots, t_m|\theta)\pi(\theta|a, b)d\theta} \\ &\approx \frac{\sum_{l=1}^{k} \theta_l f(t_1, \dots, t_m|\theta_l)}{\sum_{l=1}^{k} f(t_1, \dots, t_m|\theta_l)}. \end{aligned}$$

Usually, in a multiple hypothesis setting, the point is to identify a small proportion of interesting cases that will be investigated in detail. Then, the number of accepted hypothesis would be greater than 90% (see Efron [5]). Because of this, it seems appropriate to consider a $Beta(a, 1)$ density as the prior distribution for θ, since this prior gives a high probability to small intervals of θ close to 1. Moreover this prior includes a wide list of densities, the noninformative $Beta(1, 1)$ density among others even though we propose $a \geq 9$ to be coherent with the initial assumption that no more than 10% of null hypothesis would be declared interesting.

In this case, the posterior density is given by

$$\pi(\theta|t_1, \dots, t_m, a) \propto \theta^{a-1} \prod_{i=1}^{m} (\theta + (1-\theta)e^{n(\bar{x}_i - 1/2)}). \tag{16}$$

Table 2 Results with Method 1 with a prior density $\theta \sim Beta(a, 1)$ (for different values of a) and for simulated data from (14) with $\theta = 0.9$, different values of m and $n = 5$ observations per (each m) hypothesis

$m = 500$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.8901	0.8929	0.8942	0.8996	0.9091
$\widehat{m}_1(m_1 = 45)$	55	54	53	50	45
$prob_1$	0.6067	0.6126	0.6080	0.6097	0.6080
% Type I error	5.055	4.835	4.615	4.396	3.516
% Type II error	28.9	28.9	28.9	33.3	35.6
$m = 1000$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.8769	0.8786	0.8798	0.8826	0.8884
$\widehat{m}_1(m_1 = 107)$	123	121	120	117	112
$prob_1$	0.6635	0.6663	0.6632	0.6387	0.6273
% Type I error	4.927	4.703	4.591	4.367	3.807
% Type II error	26.168	26.168	26.168	27.103	27.103
$m = 5000$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.8984	0.8985	0.8983	0.8991	0.8999
$\widehat{m}_1(m_1 = 517)$	508	508	508	504	500
$prob_1$	0.6648	0.6649	0.6646	0.6637	0.6591
% Type I error	3.747	3.747	3.747	3.681	3.636
% Type II error	34.236	34.236	34.236	34.429	34.816
$m = 10000$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.9057	0.9057	0.9053	0.9057	0.9062
$\widehat{m}_1(m_1 = 944)$	943	943	947	943	938
$prob_1$	0.6721	0.6722	0.6735	0.6721	0.6715
% Type I error	3.379	3.379	3.412	3.379	3.335
% Type II error	32.521	32.521	32.415	32.521	32.627

A simulation from a mixture of a $N(0, 1)$ (90%) and a $N(1, 1)$ (10%) is carried out for $m = 500,\ 1000,\ 5000$ and 10000 hypothesis, with $n = 5$ observations of each hypothesis. A $Beta(a, 1)$ prior density for θ is taken where $a = 1, 7, 11, 25, 50$.

First, we use Method 1 and calculate $\widehat{\theta}, \widehat{m}_1 = m(1 - \widehat{\theta})$, – the number of rejected hypothesis with method 1–,

$$prob_1 = \widehat{P}\left(H_{(\widehat{m}_1)} = 0 | T_{(\widehat{m}_1)} = t_{(\widehat{m}_1)}\right)$$

– the highest posterior probability rejecting null hypothesis with Method 1–, and the percentage of Type I and Type II errors. The results are shown in Table 2.

Table 3 Results with Method 2 with a prior density $\theta \sim Beta(a, 1)$ (for different values of a) for the same data sets of Table 2

$m = 500$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.8901	0.8929	0.8942	0.8996	0.9091
$\widehat{i}\,(m_1 = 45)$	36	36	35	34	33
% Type I error	2.418	2.418	2.198	1.978	1.758
% Type II error	44.4	44.4	44.4	44.4	44.4
$m = 1000$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.8769	0.8786	0.8798	0.8826	0.8884
$\widehat{i}\,(m_1 = 107)$	96	95	95	94	90
% Type I error	2.352	2.352	2.352	2.352	2.239
% Type II error	29.906	30.841	30.841	31.775	34.579
$m = 5000$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.8984	0.8985	0.8983	0.8991	0.8999
$\widehat{i}\,(m_1 = 517)$	348	348	349	347	345
% Type I error	1.851	1.851	1.851	1.851	1.807
% Type II error	48.743	48.743	48.549	48.936	48.936
$m = 10000$	$a = 1$	$a = 7$	$a = 11$	$a = 25$	$a = 50$
$\widehat{\theta}$	0.9057	0.9057	0.9053	0.9057	0.9052
$\widehat{i}\,(m_1 = 944)$	654	654	656	654	652
% Type I error	1.557	1.557	1.568	1.557	1.546
% Type II error	45.657	45.657	45.551	45.657	45.763

Then, with the same data sets and $\widehat{\theta}$ from Method 1 used to compute the posterior probabilities (13), we use Method 2 and calculate $\widehat{i}$, the number of hypothesis rejected with Method 2, and the percentage of Type I and Type II errors. The results are shown in Table 3.

Observe that these two testing Bayesian procedures are robust with respect to the value of a, in the sense that a does not strongly influence the results. This new issue is good because the known Bayesian methods for testing (4) depends strongly on the parameters.

In order to compare the two proposed Bayesian methods with the FDR procedure of Benjamini and Hochberg [2], Table 4 shows, for the same data sets and $a = 11$ (an intermediate value of a), the number of null hypothesis rejected ($\widehat{n}_{BH}$, $\widehat{m}_1$, $\widehat{i}$, respectively), and the percentage of Type I and Type II errors.

Note that, with our Bayesian methods, simulations show that the number of rejected null hypotheses is more adjusted to the true than the frequentist method of Benjamini and Hochberg [2] is. For comparisons, see Table 4. In this sense the

Table 4 Results with the procedure of Benjamini and Hochberg [2] and Method 1 and Method 2 with $a = 11$, for the same data sets of Tables 2 and 3

		$\alpha = 0.05$	$\alpha = 0.05$	$\alpha = 0.05$	a = 11	a = 11	a = 11
		BH Meth.	BH Meth.	BH Meth.	Meth. 1	Meth. 1	Meth. 1
m	m_1	R_{BH}	% Type I	% Type II	$\hat{m}_1$	% Type I	% Type II
500	45	9	0	80	53	4.62	28.90
1000	107	36	0.22	68.22	120	4.59	26.17
5000	517	106	0.11	80.46	508	3.75	34.24
10000	944	187	0.03	80.51	947	3.41	32.42
		$\alpha = 0.1$	$\alpha = 0.1$	$\alpha = 0.1$	a = 11	a = 11	a = 11
		BH Meth.	BH Meth.	BH Meth.	Meth. 2	Meth. 2	Meth. 2
m	m_1	R_{BH}	% Type I	% Type II	$\hat{i}$	% Type I	% Type II
500	45	18	0.44	64.40	35	2.20	44.40
1000	107	59	0.56	49.53	95	2.35	30.84
5000	517	161	0.29	71.37	349	1.85	48.55
10000	944	307	0.23	69.70	656	1.57	45.55

procedure of Benjamini and Hochberg [2] is more conservative than each of the Bayesian methods.

One of the problems in multiple hypothesis testing with frequentist procedures is the fact that only a small number of interesting hypothesis are detected. In fact, if we want that the frequentist method achieves similar results to Method 2, simulations show that a value of $\alpha > 0.2$ is needed, but this value is not admissible.

Moreover, with our procedures the percentages of Type I errors are admissible – it does not exceed 5%–, and Method 2 is more conservative than Method 1, and the percentages of Type II errors are less than the same percentages with the frequentist procedure.

4 A Simple Hierarchical Model

Really, the parameters of the prior distribution are usually unknown and then a simple hierarchical model must be used. If we want to take a non informative prior about (a, b), Gelman et al. [6] suggests, in a different context, the use of

$$\pi(a, b) \propto (a + b)^{-5/2}. \tag{17}$$

In Sect. 2 we justified the choice of a $Beta(a, 1)$ prior to model our initial opinion about θ. Then, we propose to use $\pi(a) \propto (a + 1)^{-5/2}$. Furthermore, if we suppose (see Sect. 2) that under $H_i = 0$ the model is $N(0, 1)$ and under $H_i = 1$ is $N(1, 1)$, then the posterior density for (θ, a) when nm observations are taken is

$$\pi(\theta, a|t_1, \ldots, t_m) \propto \pi(\theta|a)\pi(a)f(t_1, \ldots, t_m|\theta)$$
$$\propto \theta^{a-1}(a+1)^{-5/2}\prod_{i=1}^{m}\left[\theta + (1-\theta)e^{n(\bar{x}_i - 1/2)}\right]$$

We use Montecarlo integration to estimate $\widehat{\theta}$. Then, we simulate a random sample $a_1, \ldots, a_h$ from $\pi(a) \propto (a + 1)^{-5/2}$ and for each a_l, a sample $\theta_1^l, \ldots, \theta_k^l$ from the prior distribution $\pi(\theta|a_l) \propto \theta^{a_l - 1}$, for $l = 1, \ldots, h$, is obtained. Finally, we estimate $\widehat{\theta}$ as:

$$\widehat{\theta} = E[\pi(\theta|t_1, \ldots, t_m, a, b)] = \frac{\int_0^\infty \int_0^1 \theta f(t_1, \ldots, t_m|\theta)\pi(\theta|a)\pi(a)d\theta da}{\int_0^\infty \int_0^1 f(t_1, \ldots, t_m|\theta)\pi(\theta|a)\pi(a)d\theta da}$$
$$\approx \frac{\sum_{l=1}^{h}\sum_{i=1}^{k}\theta_i^l f(t_1, \ldots, t_m|\theta_i^l)}{\sum_{l=1}^{k}\sum_{i=1}^{k} f(t_1, \ldots, t_m|\theta_i^l)}.$$

The same data set as in Example 3.1 is used, where a simulation from a mixture of a $N(0, 1)$ (90%) and a $N(1, 1)$ (10%) was carried out for $m = 500$, 1000, 5000 and 10000 hypothesis, with $n = 5$ observations of each hypothesis.

First, we use Method 1 to calculate $\widehat{\theta}$, $\widehat{m}_1$, the highest posterior probability of rejecting the null hypothesis and the percentage of Type I and Type II errors. The results are shown in Table 5.

Finally, we use Method 2 with the same data set, and taking $\widehat{\theta}$ from Method 1, to calculate the number of hypothesis rejected and the percentage of Type I and Type II errors. The results are shown in Table 6.

Table 5 Results using Method 1 for the hierarchical case

	$m = 500$	$m = 1000$	$m = 5000$	$m = 10000$
$\widehat{\theta}$ $(\theta = 0.9)$	0.8910	0.8781	0.8982	0.9055
$\widehat{m}_1$	55	122	509	945
$prob_1$	0.6089	0.6659	0.6655	0.6721
% Type I error	5.01	4.82	3.77	3.40
% Type II error	28.89	26.17	34.24	32.52

Table 6 Results using Method 2 for the hierarchical case

	$m = 500$	$m = 1000$	$m = 5000$	$m = 10000$
$\widehat{\theta}$ $(\theta = 0.9)$	0.8910	0.8781	0.8982	0.9055
$\widehat{m}_1$	36	95	351	655
% Type I error	2.42	2.35	1.87	1.55
% Type II error	44.44	30.84	48.36	45.55

In Sect. 3 it was shown that the results are not significantly affected by changes in the parameter a of the beta prior distribution for θ. This is the reason that, for the hierarchical case, we obtain similar results to those obtained in the previous section, because in this case the different possible values of a are replaced by the mean of its prior distribution.

5 Conclusions

The proposed methodology in this paper involves to provide a Bayesian estimator for θ (the percentage of true null hypothesis in (4)), for instance, the mean or the mode of the posterior density of θ. Posterior probabilities of $H_i = 0, i = 1, \dots, m$, are calculated and estimated by using a prior density $Beta(a, 1)$ for θ. Based on this estimator of θ, we propose two different Bayesian approaches to test (4).

Simulations show that these two Bayesian procedures are robust with respect to the value of a, in the sense that the parameter a does not strongly influence the results. This new issue is good because the known Bayesian methods for testing (4) depends strongly on the parameters.

It is well known that detecting a small number of interesting hypothesis is one of the problems in multiple hypothesis testing with frequentist approaches. In this sense, another important conclusion is that our Bayesian methods are less conservative than the procedure of Benjamini and Hochberg [2], because it allows us to reject a higher number of null hypothesis to test (4). In fact, with each of our Bayesian methods, methods 1 and 2, computations show that the number of rejected null hypotheses is more adjusted to the true than the frequentist is.

Moreover, the analyzed examples show that with our procedures the percentages of Type I errors are admissible (they do not exceed 5%), Method 2 being more conservative than Method 1, and the percentages of Type II errors are less than the same percentages with the frequentist procedure.

Acknowledgements Research supported from MEC, UCM, CM-UCM.

References

1. Barbieri M, Berger J (2004) Optimal predictive model selection. Ann Stat 32:870–897
2. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300
3. Benjamini Y, Yekutieli D (2001) The control of the false dicovery rate in multiple testing under dependency. Ann Stat 29:1165–1188
4. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Stat Sci 18(1):71–103
5. Efron B (2004) Large-scale silmultaneous hypothesis testing: the choice of the null hypothesis. J Am Stat Assoc 99(465):96–103

6. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. Chapman and Hall/CRC, London
7. Hobert J (2000) Hierarhical models: a current computational perspective. J Am Stat Assoc 95:1312–1316
8. Lehmann EL, Romano JP (2005) Testing statistical hypotheses, 3rd edn. Wiley, New York
9. Scott JG, Berger JO (2006) An exploration of aspects of Bayesian multiple testing. J Stat Plan Infer 136:2144–2162
10. Shaffer JP (1995) Multiple hypothesis testing: a review. Ann Rev Psychol 46:561–584
11. Waller RA, Duncan DB (1969) A Bayes rule for the symmetric multiple comparison problem. J Am Stat Assoc 64:1484–1503
12. Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. Am Stat 70(2):129–133
13. Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley, New York