ARTICLE TYPE: ORIGINAL RESEARCH ARTICLE

The role of the p-value in the Multitesting problem

P. Martínez-Camblor[a] and S. Pérez-Fernández[b] and S. Díaz-Coto[b]

[a] Geisel School of Medicine at Dartmouth, Hanover (NH); USA [b]Universidad de Oviedo, Asturies, Spain

ARTICLE HISTORY
Compiled October 14, 2019

ABSTRACT

Modern science frequently involves the analysis of large amount of quantitative information and the simultaneously testing of thousands or even hundreds of thousands null hypotheses. In this context, sometimes, naive deductions derived from the statistical reports substitute the rational thinking. The *reproducibility crisis* is a direct consequence of the misleading statistical conclusions. In this paper, the authors revisit some of the controversies on the implications derived from the statistical hypothesis testing. They focus on the role of the p-values on the massive multitesting problem and the loss of its standard probabilistic interpretation. The analogy between the hypothesis tests and the usual diagnostic process (both involve a decision-making) is used to point out some limitations in the probabilistic p-value interpretation and to introduce the receiver-operating characteristic, ROC, curve as a useful tool in the large-scale multitesting context. The analysis of the well-known Hedenfalk data illustrates the problem.

KEYWORDS
(Bio)markers; False-Discovery Rate; Hypothesis testing; Multitesting problem; p-value; Receiver-operating characteristic (ROC) curve.

1. Introduction

In a world dominated by uncertainty, statistical and probabilistic arguments are behind most of the scientific findings. For instance, they are crucial in the design of the experiment which led to the Bosson Higgs detection [32]. Both statistics and probability are also valuable tools for modern astronomers [17]. Moreover, the so-called *probabilistic method* [1] has been widely used by authors such as Paul Erdős (`https://en.wikipedia.org/wiki/Paul_Erd\%C5\%91s`). In medicine, the impact of the patient history on the observed measure makes that the conclusions of an overwhelming number of studies are based on statistical results. Beyond a corporativist claim, the presence of statistical professionals in most of the research projects is becoming a reality. However, probabilistic laws are frequently different to the logical ones; for instance, the assumptions i) $\mathcal{P}(A > B) > 1/2$ and ii) $\mathcal{P}(B > C) > 1/2$ do not imply that $\mathcal{P}(A > C) > 1/2$ (probability does not satisfy the transitivity property). Correct statistical results are sometimes misunderstood and therefore the con-

CONTACT: Pablo Martínez-Camblor. Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, 7 Lebanon Street, Suite 309, Hinman Box 7251, Hanover, NH 03755, USA. Email: Pablo.Martinez.Camblor@dartmouth.edu

clusions derived from these misconceptions are wrong as well. Biomedical researchers frequently simplify their findings and translate them in a *yes/no* response. This simplification is especially dangerous when it is based on statistical hypothesis tests and on the omnipresent significance level $\alpha = 0.05$. Moreover, in the last decades, technological advances have drastically increased the available information. Modern science frequently produces data on thousands of different characteristics (variables). The -omic technologies (genomics, transcriptomics, proteomics, etc.) stand for the most relevant examples although other fields like brain imaging or spatial epidemiology have also increased substantially the size of the collected data. A usual practice is to measure all this information in two or more groups of individuals (sample units) and, by using one particular statistical test, determining whether, in those groups, the behavior of the collected variables is different or not. In this context, individual p-values lose any probabilistic interpretation and classical multitesting correction procedures (the Bonferroni is the most popular) lead to drastic increments in the Type II error. In order to deal with this problem, more liberal criteria have been proposed in the specialized literature (see, for instance, Farcomeni [16] and references therein), being the False-Discovery Rate (FDR) [26] the most popular.

In this manuscript, the authors ponder on the role of the p-values in the multiple hypothesis testing context. Some of the interpretations and limitations of the usual multitesting procedures are pointed out. We try to avoid excessively technical expressions in order to bring this discussion near to statistical practitioners although we realize that for a full understanding of the manuscript, some statistical background in multitesting problems is required; notice that most of the references dealing with the multitesting problem are published in strongly theoretical journals. The remainder of the paper is structured as follows. Although it is not the main focus of the manuscript, the concept and implications of individual p-values are discussed in Section 2. Section 3 is devoted to the controversy over whether p-values should be adjusted or not in the multitesting context. The well-known False-Discovery Rate (FDR) and some of its main competitors are considered in Section 4. In Section 5, we introduce the main idea of this paper; in the multitesting context, we propose to interpret the p-value as a marker of the null hypothesis being true. That is, for each subject (read hypothesis), small p-values would be associated, and just associated, with larger likelihood of being positive (null hypothesis untrue). This interpretation allows to point out some of the limitations of the multitesting procedures and helps us to a better understanding of the obtained results. In addition, the well-known receiver-operating characteristic (ROC) curve [19] appears as an useful tool to be considered in the multitesting problem. A study of the well-known Hedenfalk [20] data is displayed in Section 6. Finally, in Section 7 we present our main conclusions.

2. The elusive concept of p-value

From a mathematical point of view, let $H_0 \ : \Theta = 0$ ($H_1 \ : \Theta \neq 0$) be a usual generic statistical hypothesis testing and $\widehat{\Theta}_X$ the *evidence* resulting from the sample realization, X. The p-value is defined by $p = 1 - \mathcal{P}(\Theta \in (-|\widehat{\Theta}_X|, |\widehat{\Theta}_X|)\,|H_0)$ i.e., the probability that the obtained result, or one more divergent from the null, was observed given that the null hypothesis is true. Since the nominal level α ($= 0.05$, usually) is fixed, the null rejection ($p < \alpha$) can imply not only that H_0 is false but also that the sample belongs to an unlikely *family (subset) under the null.* That is, the common conclusion derived of rejecting the null based on a low p-value contains the risk of rejecting a true null: this possibility is known as Type I error and its probability is controlled by α.

Criticisms on the real implication of the p-value are not new (see, for instance, Cohen [11]) but this controversy still produces references (see, for instance, Wellek [36] and the subsequent comments to this paper). In 2015, the journal *Basic and Applied Social Psychology* published an Editorial banning the use of p-values [31]. Polemic continued in the Editorial of the journal Significance [30]. In 2016, The American Statistical Association published a statement [35] with six principles regarding the *p-values* use which tries *to improve the conduct and interpretation of quantitative science.*

Another negative feature of the p-values is that they strongly depend on the sample size. They decrease when the sample size increases. Demidenko [13] suggested using an effect size measure, as many statisticians already do, but expressed on the probability scale. Particularly, he proposed to use the so-called D-value which, for the two sample-problem and if X and Y stand for the random variables from which the two independent samples were drawn, is defined by $\mathcal{P}(X < Y)$. Perhaps, the lack of the transitivity property is its most relevant handicap.

3. The multitesting problem: to adjust or not to adjust, that's the question

The multitesting problem appears when two or more hypotheses tests are considered simultaneously, i.e. we have $H_{0,i} \ : \ \Theta_i = 0$ ($H_{1,i} \ : \ \Theta_i \neq 0$) with $1 \leq i \leq N$. In this context, considering the nominal level as in the single hypothesis setting, the probability of committing any (false positives) increases. Adjusting the significance level allows to control the probability of Type I error at the original fixed level. There exist several procedures for adjusting the p-values, among all of them, the Bonferroni method [7] is the most popular. In biomedicine, the use of these adjustments -however- is still controversial.

Averse arguments to the multiplicity adjustment are mainly that i) when one adjusts, the null hypothesis actually contrasted is the intersection of the N originally involved nulls; this (new) null is, in general, lack of interest [24], ii) adjustment implies an unnecessary increment in the probability of committing a Type II error and iii) it is not clear what number of tests should be considered in the adjustment: number of tests done, number of tests included in the document, other studies which deal with the same research question, etc. The implication of this last point is particularly interesting. Because the obtained p-value should be modified if, in the future, new studies dealing with the same research question were conducted (or just published?); the study conclusion would be submitted to constant potential changes [23]. From a technical point of view, it is clear that if N true-null tests are simultaneously performed at a fixed nominal level α, the probability that some of the (N) obtained p-values were significant (below α) is greater than α (in fact, assuming independence among the N tests, it is $1 - (1 - \alpha)^N$, almost 1 for $N \geq 100$). Bender and Lange [3] argued that the multiplicity adjustment is not just related with the *intersection hypothesis* but allows to derive more plausible conclusions regarding to all the hypotheses considered in the study. There is not an easy answer to the question of when the adjustments for multiple tests are necessary [4]. In confirmatory studies with pre-fixed goals such as clinical trials, in which the hypothesis testing is used as a tool for determining the final decision, the multiplicity adjustment is mandatory [25]. In observational exploratory studies, the adjustment is not a requirement [34] although obtained results should be carefully interpreted.

The multitesting problem changes in the massive data context. In these analyses, the previous hypotheses are frequently related to *the existence* of a number of *effects* (variables behaving different between positive and negative subjects) among a huge number of potential ones. The effects are often selected among those with the smallest p-values. In this context, to control the so-called family wise error rate (FWER), that is, to control the probability of committing any Type I error (one non-effect declared as effect) is too restrictive because it provokes a drastically increment in the Type II error probability. Besides, most of those procedures assume independence among the hypotheses to test making them even more conservative. Conversely, making no control on the fixed nominal level supposes a non-assumable number of spurious (erroneous) rejections. In order to deal with this question, more liberal criteria have been proposed. In the next section we revise some of them.

4. The False-Discovery Rate and its competitors

The FDR was studied, popularized and formally defined by Benjamini and Hochberg [5] in one of the most cited statistical papers in the history. Previously, Seeger [26] elaborate and discussed a stepwise multiple testing procedure for controlling the proportion of false discoveries among all discoveries and the same algorithm was considered by Simes [28] who proved that it controls FWER when all hypotheses are true. FDR s defined as the expected proportion of (false) spurious effects declared i.e., the expected value for the false-discovery proportion, FDP, that is,

$$\begin{aligned} \text{FDR} =&\mathbb{E}[V/(R \vee 1)] \\ =&\mathbb{E}\left[V/R \,|\, R > 0\right] \cdot \mathcal{P}(R > 0), \end{aligned} \tag{1}$$

where V is the random variable modelling the number of true nulls (erroneously) rejected and $(R \vee 1)$ stands for the maximum between the total number of rejections, R, and 1, according to some significance rule. Notice that, with this notation, the FWER tries to control $\mathcal{P}(V > 0)$. The FDR is a frequentist well established definition for the multiple hypothesis testing errors. Besides, it has some interesting properties: it is equivalent to the FWER when all the null hypotheses tested, N, are true and is more liberal than the FWER criterion otherwise. The second point is that it allows to increase considerably the number of rejections and, likely, the number of detected effects. Notice that, in a number of large-scale experiments (for instance in microarray genome-wide scans), it is usually supposed that most of the nulls are true. The usual goal of these studies is to identify a subset of interesting factors for future confirmatory experiments and the relative role of Type I and Type II errors can change with respect to the (single) hypothesis testing framework.

Figure 1 depicts, visually, the different philosophy of both the FWER and the FDR criteria. While the FWER tries to commit no Type I error and, in fact, just a small percentage of the studies should contain rejected true-nulls, the FDR assumes some percentage of false discoveries in each study.

Other criteria have been proposed in the specialized literature including the positive false discovery rate (pFDR) suggested by Storey [29]. Most recent works deal with the problem of controlling the tail probability of false positive discoveries. For instance, Genovese and Wasserman [18] proposed to control the tail probability of the false discovery proportion (FDP) by the so-called false discovery exceedance (FDX); that is, to control $\mathcal{P}(V/(R \vee 1) > \alpha)$. Similarly, Dudoit et al. [14] suggested controlling the generalized family-wise error rate (gFWER) defined as the probability of committing more than k Type I errors i.e., $\mathcal{P}(V > k)$. Carvajal-Rodríguez et al. [8] proposed the sequential goodness of fit (SGoF) strategy which rejects the number of null hypotheses equal to the difference between the observed and the expected amounts of p-values below a given threshold under the assumption that all nulls are true. The SGoF procedure have been extended by Castro-Conde et al. [9]. In general, new references studying, extending and proposing new multitesting procedures are continuously published in the specialized literature.

Assuming independence among the N involved tests, the p-values cumulative distribution function is given by the mixture

$$G(t) = \mathcal{P}(p \leq t) = (1 - \pi) \cdot F_0(t) + \pi \cdot F_1(t) \qquad t \in [0, 1], \tag{2}$$

where F_0 stands for the distribution function of the p-values when the null hypothesis is true (we assume here that $F_0(t) = t$ with $t \in [0, 1]$, i.e., we assume that, under the null, each single p-value is uniformly distributed on $[0, 1]$). This is a reasonable proviso, although it is only true if the null is simple and the distribution, under the null, of the used test statistic is continuous and known; the true p-value distribution is only stochastically dominated by the uniform if its distribution is discrete or the p-value is estimated by a resampling method [16], F_1 is the CDF for the p-values when the null is false and π is the proportion of false nulls (effects) in the N considered tests; i.e., $\pi = N_1/N$ with N_1 the number of false nulls. That is, π is the *prevalence* of effects on the set of consider hypotheses. Hence, fixed a threshold, t (the null is rejected if the p-value is below t), the false discovery proportion (FDP) is defined by,

$$\mathrm{FDP}(t) = \frac{\sum_{i \in \mathcal{I}_0} I(p_i < t)}{\sum_{i=1}^{N} I(p_i < t) \vee 1} = v_t \,/ (r_t \vee 1) \qquad t \in [0, 1], \tag{3}$$

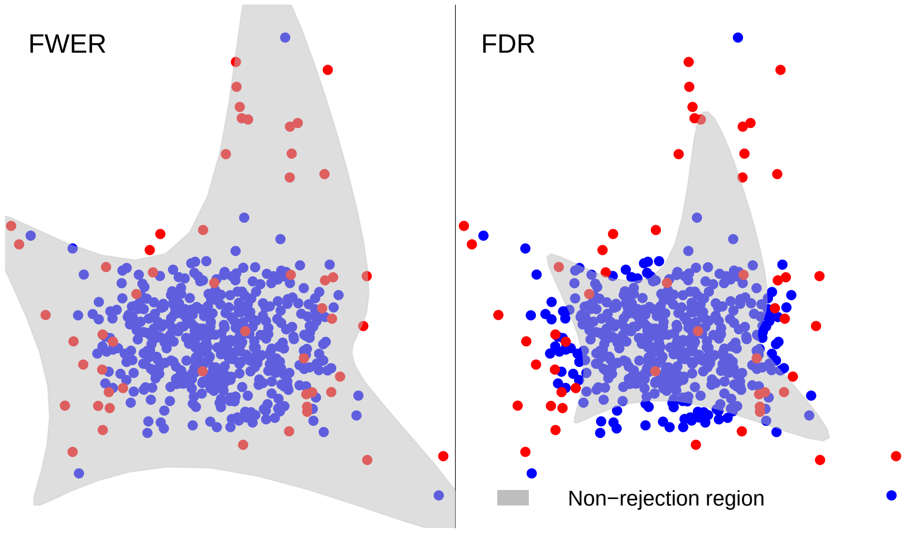

Figure 1. Blue-dots stand the true-nulls, red-dots the false-nulls. Dots outside the gray region are rejected.

with $\mathcal{I}_0$ the set of indices for the true nulls ($\#\mathcal{I}_0 = N - N_1$) and $I(A)$ the standard indicator function (takes a value of 1 if A is true and 0 otherwise). Conditioning to the number of rejected hypothesis is r_t, under the mixture model, v_t follows a binomial distribution with number of repetitions r_t and success probability $(1-\pi)\cdot t\,/G(t)$. Besides r_t follows a binomial distribution with N the number of repetitions and probability of success $G(t)$. Then

$$\begin{aligned}
\mathrm{FDR}(t) =&\mathbb{E}[\mathrm{FDP}(t)]\\
=&\mathbb{E}[\mathbb{E}[v_t/(r_t \vee 1)|r_t]]\\
=&\sum_{r_t=1}^{N} \frac{1}{r_t}\frac{r_t(1-\pi)t}{G(t)}\binom{N}{r_t} G(t)^{r_t}(1-G(t))^{N-r_t}\\
=&\frac{(1-\pi)t}{G(t)}(1-(1-G(t))^N) = \frac{(1-\pi)t}{G(t)} + \mathcal{O}((1-G(t))^N). \qquad (4)
\end{aligned}$$

Due to the current state-of-the-art (N is usually a really big number), the error term can be removed. Therefore, given a sorted sample of p-values, $\{p_{(1)}, p_{(2)}, \cdots, p_{(N)}\}$, a simple plug-in method determines that, for each $i \in \{1, \cdots, N\}$,

$$\widehat{\mathrm{FDR}}(p_{(i)}) = \frac{(1-\pi)\cdot p_{(i)}}{\widehat{G_N}(p_{(i)})} \leq \frac{N \cdot p_{(i)}}{i}.$$

where $\widehat{G_N}$ is the empirical cumulative distribution function of G. For a fixed nominal level, α, Benjamini and Hochberg [5] (BH) proposed to select the threshold by $t_{BH} = \max\{p_{(j)} : p_{(j)} \leq j \cdot \alpha/N,\ 1 \leq j \leq N\}$, and then

$$\widehat{\mathrm{FDR}}(t_{BH}) = \frac{(1-\pi)\cdot t_{BH}}{\widehat{G_N}(t_{BH})} \leq \alpha,$$

so, the expected value of false discovery proportion is controlled at the desired level α. Notice that i) the properties of the p-value are used for controlling the FDR but the FDR has not these properties, ii) a less conservative approximation could be developed by making any previous estimation of the parameter π (see, for instance, Dalmasso, Broët and Moreau [12]) and iii) although the proof of the previous results has been extended to certain types of dependencies among the hypotheses tested [6], it is strongly dependent on the relationships among them.

5. p-Value as marker. The prevalence matters

It seems clear that, in the massive-data context, individual p-values lose their probabilistic interpretation. Therefore, and due to the p-values stochastic nature, it can be assumed that we have a measure which has different behavior in those subjects (hypotheses) belonging to the positive group (false-null) and in those which belong to the negative group (true-null). Hence, p-values can be seen as a marker of the likelihood that a subject is or is not positive. Of course, smallest p-values are associated with highest probabilities of being within the positive group. Left panel of Figure 2 represents the probability density functions (PDF) of the z-values ($\Phi^{-1}(p)$ with p the p-value and Φ the cumulative distribution function (CDF) of the standard normal distribution) under both the null (blue line) and the alternative (red line) hypotheses. Each null hypothesis, $H_{0,i} : \mu_{1,i} = \mu_{2,i}$ $(1 \leq i \leq N)$, was checked using the Student T-test. We consider standard normal distributed samples and alternatives detected with a statistical power of 0.8 (at $\alpha = 0.05$). The first lesson to learn is that, as it is well-known in the diagnostic context, the prevalence is a crucial parameter in order to determine that a subject, $\mathcal{H}$, with an abnormal marker value is really a positive subject. Notice that, if the marker value for the subject $\mathcal{H}$ is τ, then

$$\begin{aligned}\mathcal{P}(\mathcal{H} \in H_1 | T \leq \tau) =& \frac{\pi \cdot \mathcal{P}(T \leq \tau \,|\, \mathcal{H} \in H_1)}{(1-\pi) \cdot \mathcal{P}(T \leq \tau \,|\, \mathcal{H} \in H_0) + \pi \cdot \mathcal{P}(T \leq \tau \,|\, \mathcal{H} \in H_1)} \\ =& \frac{\pi \cdot S_e(\tau)}{(1-\pi) \cdot (1 - S_p(\tau)) + \pi \cdot S_e(\tau)},\end{aligned}$$

where H_1 and H_0 stand for the positive and negative groups, respectively, T models the marker behavior, π is the prevalence of the studied characteristic in the population, S_e the sensitivity (that is, the probability that a positive subject is correctly classified as positive) and S_p the specificity (the probability that a negative subject is correctly classified as negative). Even using a classification rule with both sensitivity and specificity of 0.95, if the prevalence is 0.003, the probability that a random selected subject classified as positive was really a positive is 0.05 (from the above equation, $0.003 \cdot 0.95/(0.997 \cdot 0.95 + 0.003 \cdot 0.95)$). At right, Figure 2 depicts the above-mentioned probability against the real proportion of false-null for diagnostic procedures with different sensitivities/specificities. Notice that, when the proportion of false-null, called here *prevalence* (N_1/N) is zero, the probability of success in the positive group is mandatorily zero. Besides, since in practice the number of effects in the set are expected to be small [15], we report the prevalence in log-scale for highlighting the smaller values.

The point is that, given a random vector modelling the p-values behavior for the

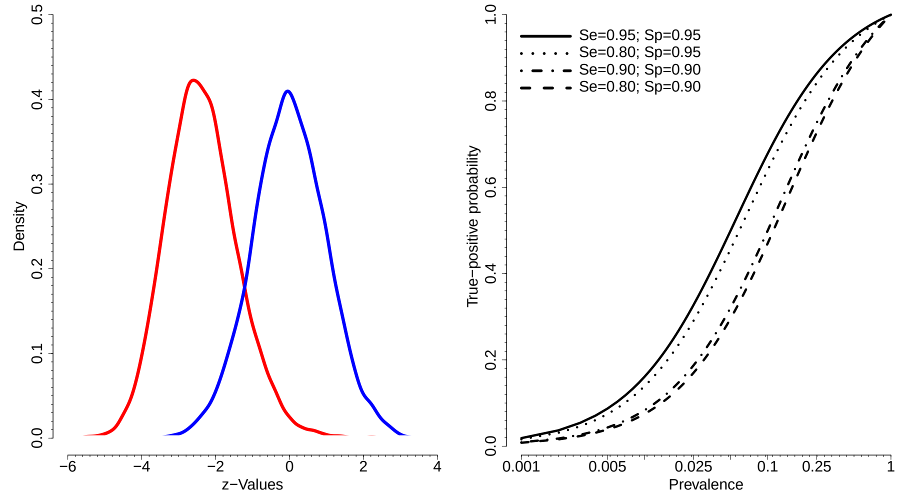

Figure 2. Left, probability density functions (PDF) of the z-values ($\Phi^{-1}(p)$ with p the p-values and Φ the cumulative distribution function (CDF) of the standard normal distribution) under both the null (blue line) and the alternative (red line) hypotheses. Right, probability that a subject classified as positive was really positive against the prevalence (in log-scale) for diagnostic procedures with different sensitivity (S_e) and specificity (S_p) values.

false-null (positive) group, $\boldsymbol{\xi} = \{\xi_1, \cdots, \xi_{N_1}\}$, the distribution of the k-th sorted statistic is,

$$\mathcal{P}\{\xi_{(k)} \leq t\} = F_{\xi_{(k)}}(t) = \sum_{j=k}^{N_1} \binom{N_1}{j} F_1(t)^j (1 - F_1(t))^{N_1 - j}.$$

We simultaneously check these hypotheses joint with another N_0 true-nulls. Let $\boldsymbol{\chi} = \{\chi_1, \cdots, \chi_{N_0}\}$ be the random vector modelling the p-values behavior for the true-null group. In this case, the distribution of the l-th sorted statistics is,

$$\mathcal{P}\{\chi_{(l)} \leq t\} = F_{\chi_{(l)}}(t) = \sum_{i=l}^{N_0} \binom{N_0}{i} t^i (1 - t)^{N_0 - i}.$$

Rejecting k false-null hypotheses implies rejecting as well those true-null with p-values below $\xi_{(k)}$ and we know that,

$$\mathcal{P}\{\chi_{(l)} \leq \xi_{(k)}\} = \int F_{\xi_{(k)}}(t) dF_{\chi_{(l)}}(t).$$

This implies that the threshold required for detecting a particular percentage of false-null is strongly affected by the number of true-null hypotheses which are included and simultaneously tested. We consider a fixed number of false-null, $N_1 = 100$, with Type II error probability of 0.2 for a Type I error probability of 0.05. Figure 3, at left, depicts the FDR required for detecting 10, 25, 50, and 80 effects (false-null) against the prevalence in log-scale ($\pi = N_1/(N_0 + N_1)$). For the smallest considered prevalence, we really do not have the ability to detect most of the effects and the probability of rejecting a true-null is larger than the probability of rejecting a false-null. At right, Figure 3 shows the FDR against the threshold (in log-scale) for different prevalence. The FDR has a clear inflexion point which strongly depends on the prevalence. Besides, the FDR has a rapid increment with the threshold for lowest prevalence.

5.1. Dependency structures

As we already have mentioned, the validity of most of the adjusting procedures, are proved assuming the independence among the N considered tests, the so-called *independence assumption.* Although this assumption is reasonable in a number of practical situations (see Clarke and Hall [10]), it may be a source of serious drawbacks. The effect of the potential relationship between the obtained p-values has been previously considered in the literature and there exists a number of papers dealing with it (see, for instance, Martinez-Camblor [22] and references therein). However, many of them assume hypotheses which are only reasonable for particular settings. In some fields, for instance in epigenomic or in mRNA analysis, it is known the strong relationship among the considered tests. This relationship frequently provokes asymmetry in the distribution of the observed number of rejections under the null. This asymmetry implies that to control a expected value of false discoveries (FDR) does not imply to have a good control of the number of false discoveries in a particular sample realization. Besides, the idea of having N subjects disappears, and we return to the undesired scenario in which we have a N-dimensional random sample with size 1. Figure 4 de-

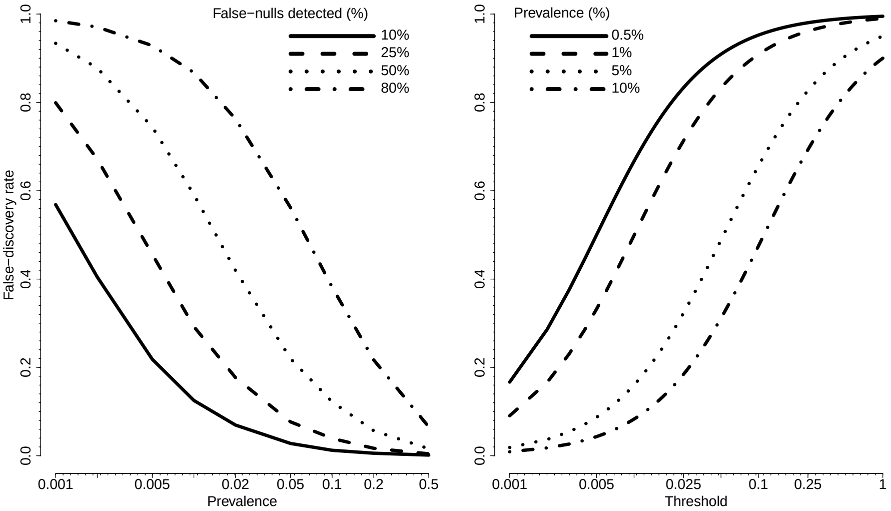

Figure 3. Left, FDR for detecting a fixed number of false-null hypotheses (10, 25, 50 and 80 are considered) against prevalence (log-scale) for a problem in which $N_1 = 100$ and power is 0.8 at $\alpha = 0.05$. Normal samples and T-tests are considered. Right, FDR against threshold (log-scale) for the same problem and different prevalence.

picts the PDF of the observed number of rejections ($\alpha = 0.05$, 0.10) for simulated situations in which all null hypotheses tested are true. Correlation structures are extracted from two real-world studies: the Hedenfalk data [20] (3226 genes directly downloaded from `http://faculty.washington.edu/~jstorey/qvalues/results.html`) and the Shen data [27] (the dataset are publicly available at the Gene Expression Omnibus (GEO) page, with accession number GSE37988, `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37988`. A total of 24977 autosomal CpG sites are considered). The asymmetry increases the probability of having an abnormal large number of observed rejections.

Although it was proved that, for some correlation structures, the BH procedure still controls the FDR [6], it is worth remarking that the FDR is an *expected value* and then, it is not the most appropriate summary measure for strongly asymmetric distributions. In these cases, to control the FDR does not avoid that the probability of observing large values of FDPs is still high even when all tested hypotheses are true.

6. Real-world application: the *Hedenfalk* data

The Hedenfalk data [20] stands for an illustrative and well-known example which, in the present framework, has been previously considered by several authors. One of the goals of this microarray study was to find genes differentially expressed in breast cancer patients with mutations in the BRCA1 gene relative to those with BRCA2 mutation. A total of 3226 ($= N$) genes were studied on 7 patients with BRCA1- and on 8 with BRCA2-mutation. The mean±standard deviation for the p- and z-values obtained from the 3226 two-side Student tests (one for each gene) on the Hedenfalk data (without any data pre-processing), were 0.372 ± 0.30 and -0.471 ± 1.15, respectively. Figure 5 depicts the histogram for the 3226 p-values (top-left) and the kernel density estimation for the respective z-values (top-right). At level $\alpha = 0.10$, the BH procedure declares 124 effects, those below 0.00385 ($= \tau_0$).

A total of 840 out of 3226 p-values, 26%, are below 0.1. Notice that, based on the left

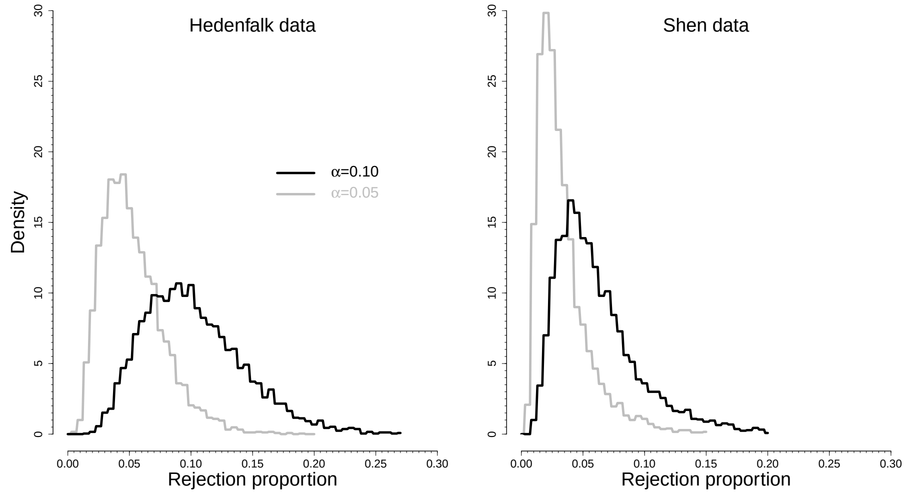

Figure 4. Distribution of the number of observed rejections at levels $\alpha = 0.05$ (gray) and 0.10 (black) based on simulations (5000 iterations) in which all the nulls are true. Considered correlations structures are extracted from the Hedenfalk data (3226 genes), left, and from the Shen data (24977 CpG sites), right.

panel of Figure 4 (black line) the probability of observing this value when all the null hypotheses were true is zero. Assuming independence among the hypotheses, and that, for an specificity of 0.1, the sensitivity of the test is 0.8 (Figure 5, bottom-left, depicts the resulting ROC curve), we have that $F_1(\tau_0) = \mathcal{P}\{p < \tau_0|H_1\} = 0.2935$ (computed similarly to the panel right in Figure 2). Since $\widehat{G}_N(\tau_0) = 124/3226$, in this scenario, $\tilde{\pi}_0 = 0.1195$. With this prevalence, the expected proportion of false-null hypotheses detected is 0.291 (112/385). From equation (4), using the prevalence estimation, we can obtain a new bound for the FDR. Particularly, we detect 147 effects with a threshold of 0.0051 ($= \tau_1$) and expected proportion of false-null hypotheses detected of 0.329 ($\tilde{\pi}_1 = 0.1243$). A new iteration reports 148 declared effects ($\tau_2 = 0.0052$) and a prevalence of 0.1253 ($= \tilde{\pi}_2$) which means a similar expected proportion of false-null hypotheses detected of 0.329. The fourth iteration reports the same results. Figure 5, bottom-right, represent the iterative process.

7. Discussion

Statistical procedures are frequently performed routinely and without the previous checking of the required assumptions. The derived conclusions are sometimes misunderstood and are not taken with the appropriate caution. Those actions contribute to the *reproducibility crisis* and to the deterioration of the sciences credibility [2]. The problem gets worse when the studies involves thousands of statistical hypotheses which cannot be handled individually nor carefully. Such is the case of most of the studies in the so-called -omic sciences in which, commonly, thousands or even hundreds of thousands of null hypotheses are simultaneously tested and, once a threshold is computed, a subset of them are considered as *effects*. But, of course, statistical analysis cannot replace the rational thinking and the derived conclusions should be carefully considered [35]. Knowing the real implications of the selected threshold and the risks (limitations) of the decisions based on statistical hypothesis testing is crucial to get a good understanding of the observed results.

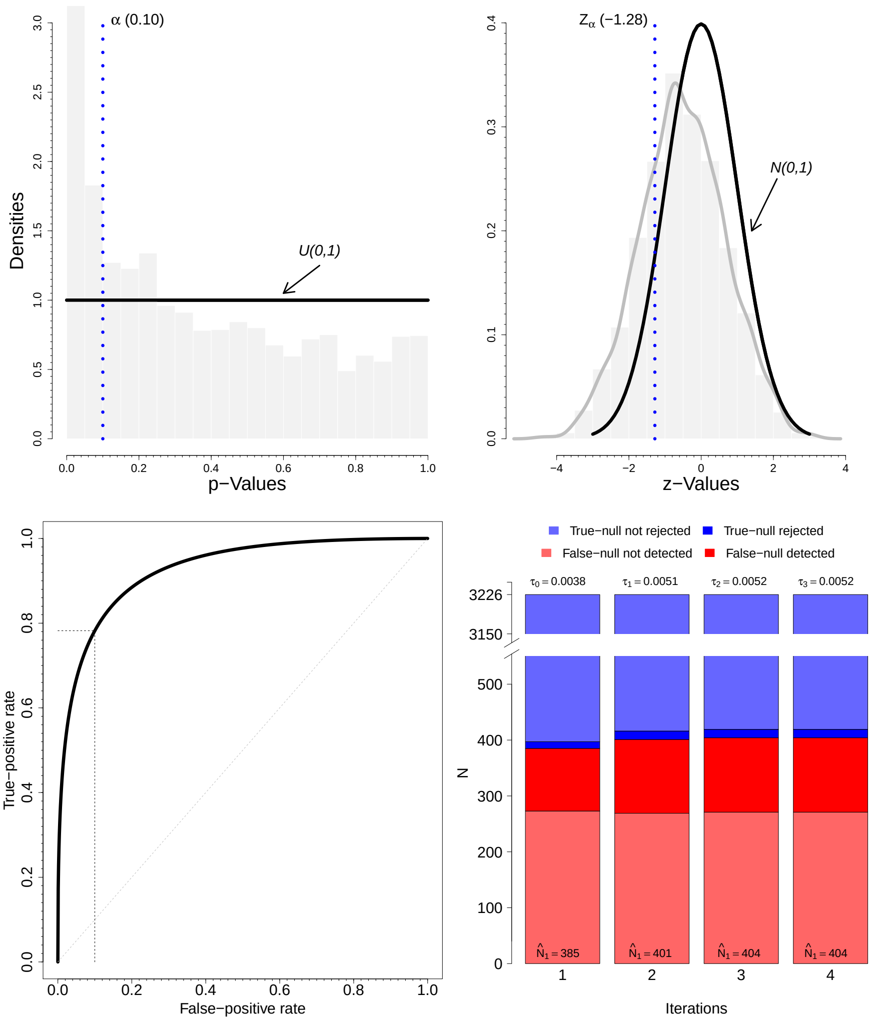

Figure 5. Top, p- (left) and z-values (right) histograms for the 3226 genes of the Hedenfalk data. Bottom, ROC curve for the problem assuming the alternative hypotheses are detected with a power of 0.8 at level 0.10 (left). The panel in the bottom-right stands for the graphical representation of the iterative method proposed for estimating the FDR and the percentage of false-null hypotheses detected.

In this paper, we revisited some relevant controversies regarding to the interpretation and the limitations of the statistical hypothesis tests in the multitesting context. Avoiding the usual (perhaps) excessively probabilistic language, we review the false discovery rate (FDR) criteria and point out some of its limitations. We include some simple equations which allow to understand the algorithm popularized by Benjamini and Hochberg [5]. These are helpful for seeing that, in this procedure, the p-value is just a mean for getting the objective of controlling the FDR. We also insist in the relevancy of the ratio of false- and true-null hypotheses in order to know what the real capacity of our study for detecting the real existing effects is [21]. The analogies between the massive data testing and the diagnostic problem allow a better understanding of the prevalence (proportion of false-null hypotheses) relevancy and its crucial role for computing the probability that a rejected null hypothesis is actually an effect. Although, of course, the contexts are not equal, the clear link allows to take advantage of some tools used in the diagnostic process context, such as the receiver-operating characteristic (ROC) curve, for a better depicting and understanding of the considered scenario. Arguing similarly to the problem of sample size compute, we fix some conditions on the p-value distribution under the alternative hypotheses to provide prevalence estimations. These estimations help us to know more about the problem we are dealing with. Important to remark that, due to the stochastic nature of the p-value behavior, to know the real number of false-null hypotheses does not imply that we can detect exactly which are those false-null hypotheses. Besides, remarkable that smallest p-value does not imply strongest effects [33].

Our results show clearly how the noise produced by the true-null hypotheses included in the analysis provokes a relevant lost in our capacity to detect actual false-null hypotheses. Therefore, in the large-scale testing context, it is advisable to reduce the number of tests performed by using previous information about the region in which the searched effects can be located. In addition, the presence of not negligible correlation between the performed tests complicates the problem. The role of the so-called independence assumption should be considered in order to interpret the obtained ROC curves. Simulation studies report additional information on the false-discovery proportion (FDP) which enlighten the complexity of the task we are dealing with.

In short, not each obtained conclusion can be reduced to a simple yes/no answer. Sometimes, the learning derived from scientific studies can be ambiguous and/or inconclusive. To keep this tone in the written sentences can be less attractive but more realistic about what the real state of the art is. In the massive data analysis context, to provide some information about the *effects* prevalence in the total setting of hypotheses performed can help us to settle how realistic the derived conclusions are.

Acknowledgement

This work is supported by the Grants MTM2015-63971-P, MTM2014-55966-P and MTM2017-89422-P (ERDF support included) from the Spanish Ministerio of Economía y Competitividad, FC-15-GRUPIN14-101 from the Asturies Government and Severo Ochoa Grant BP16118 (this one for S. Pérez-Fernández).

References

[1] N. Alon and J. Spencer, *The probabilistic method*, 4th ed., Wiley Publishing, 2016.

[2] M. Baker, *1,500 scientists lift the lid on reproducibility*, Nature 553 (2016), pp. 452–454.

[3] R. Bender and S. Lange, *What's wrong with arguments against multiplicity adjustments*, Letter to the editor concerning BMJ 316 (1998), pp. 1236–1238.

[4] R. Bender and S. Lange, *Adjusting for multiple testing-when and how?*, Journal of Clinical Epidemiology 54 (2001), pp. 343 – 349.

[5] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological) 57 (1995), pp. 289–300.

[6] Y. Benjamini and D. Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, The Annals of Statistics 29 (2001), pp. 1165–1188.

[7] J. Bland and D. Altman, *Multiple significance tests: the Bonferroni method*, BMJ 310 (1995), p. 170.

[8] A. Carvajal-Rodríguez, J. de Uña-Alvarez, and E. Rolán-Alvarez, *A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests*, BMC Bioinformatics 10 (2009), pp. 209–220.

[9] I. Castro-Conde, S. Dohler, and J. de Uña Álvarez, *An extended sequential goodness-of-fit multiple testing method for discrete data*, Statistical Methods in Medical Research 26 (2017), pp. 2356–2375.

[10] S. Clarke and P. Hall, *Robustness of multiple testing procedures against dependence*, The Annals of Statistics 37 (2009), pp. 332–358.

[11] J. Cohen, *The earth is round ($p < .05$)*, American Psychologist 49 (1994), pp. 997–1003.

[12] C. Dalmasso, P. Broët, and T. Moreau, *A simple procedure for estimating the false discovery rate*, Bioinformatics 21 (2005), pp. 660–668.

[13] E. Demidenko, *The p-value you can't buy*, The American Statistician 70 (2016), pp. 33–38.

[14] S. Dudoit, M. van der Laan, and M. Birkner, *Multiple testing procedure for controlling tail probability error rates*, Technical Report, 166, Divisionof Biostatistics, California University, Berkeley. (2004).

[15] B. Efron, *Correlation and large-scale simultaneous significance testing*, Journal of the American Statistical Association 102 (2007), pp. 93–103.

[16] A. Farcomeni, *A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion*, Statistical Methods in Medical Research 17 (2008), pp. 347–388.

[17] E. Feigelson and G. Babu, *Statistical methods for Astronomy*, in *Planets, stars and stellar systems: Volume 2: Astronomical techniques, software, and data*, T.D. Oswalt and H.E. Bond, eds., Springer Netherlands, Dordrecht (2013), pp. 445–480.

[18] C. Genovese and L. Wasserman, *Exceedance control of the false discovery proportion*, Journal of the American Statistical Association 101 (2006), pp. 1408–1417.

[19] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*, Wiley, New York, 1966.

[20] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. Kallioniemi, B. Wilfond, A. Borg, J. Trent, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, and G. Sauter, *Gene-expression profiles in hereditary breast cancer*, New England Journal of Medicine 344 (2001), pp. 539–548.

[21] J. Ioannidis, *Why most published research findings are false*, PLoS Medicine 2 (2005), p. e124.

[22] P. Martínez-Camblor, *On correlated z-values distribution in hypothesis testing*, Computational Statistics & Data Analysis 79 (2014), pp. 30–43.

[23] T. Perneger, *What's wrong with Bonferroni adjustments*, BMJ 316 (1998), pp. 1236–1238.

[24] K. Rothman, *No adjustments are needed for multiple comparisons*, Epidemiology 1 (1990), pp. 43–46.

[25] A. Sankoh, M. Huque, and S. Dubey, *Some comments on frequently used multiple endpoint adjustment methods in clinical trials*, Statistics in Medicine 16 (1997), pp. 2529–2542.

[26] P. Seeger, *A note on a method for the analysis of significances en masse*, Technometrics 10 (1968), pp. 586–593.

[27] J. Shen, S. Wang, Y. Zhang, M. Kappil, H. Wu, M. Kibriya, Q. Wang, F. Jasmine, H. Ahsan, P. Lee, M. Yu, C. Chen, and R. Santella, *Genome-wide DNA methylation profiles in hepatocellular carcinoma*, Hepatology 55 (2012), pp. 1799–1808.

[28] R. Simes, *An improved Bonferroni procedure for multiple tests of significance*, Biometrika 73 (1986), pp. 751–754.

[29] J. Storey, *The positive false discovery rate: a Bayesian interpretation and the q-value*, The Annals of Statistics 31 (2003), pp. 2013–2035.

[30] B. Tarran and M. Wininger, *Editorial: A psychology journal bans p-values*, Significance 12 (2015), pp. 2–7.

[31] D. Trafimow and M. Marks, *Editorial*, Basic and Applied Social Psychology 37 (2015), pp. 1–2.

[32] D. van Dyk, *The role of statistics in the discovery of a Higgs Boson*, Annual Review of Statistics and Its Application 1 (2014), pp. 41–59.

[33] A. Vexler, J. Yu, Y. Zhao, A. Hutson, and G. Gurevich, *Expected p-values in light of an ROC curve analysis applied to optimal multiple testing procedures*, Statistical Methods in Medical Research 0 (2017), p. 0962280217704451.

[34] E. von Elm, D. Altman, M. Egger, S. Pocock, P. Gøtzsche, J. Vandenbroucke, and for the STROBE Initiative, *The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies*, PLOS Medicine 4 (2007), pp. 1–5.

[35] R. Wasserstein and N. Lazar, *The ASA's statement on p-values: context, process, and purpose*, The American Statistician 70 (2016), pp. 129–133.

[36] S. Wellek, *A critical evaluation of the current p-value controversy*, Biometrical Journal 1 (2017), pp. 1–19.