The problem of false discoveries from a metrological point of view

Giampaolo E D'Errico

Istituto Nazionale di Ricerca Metrologica (INRIM), Torino (I)

E-mail: g.derrico@inrim.it

Abstract. Valid inference drawn from analysis of experimental results needs scientific grounds, whereas conclusions based on statistical significance tests or hypothesis testing may be problematic, especially when dealing with a multiplicity of tested hypotheses, as in experiments performed on bio-molecules. The problem of false discovery rate is focused in the present paper, aiming at eliciting application of sound criteria for rejection/acceptance of hypotheses and related methods of uncertainty characterisation.

1. Introduction

The qualifier 'statistically significant' attributed to an experimental result is not the same as 'scientifically meaningful'. It is well known that 'significant' is one of the most disputed definitions in scientific literature (see e.g. [1]), and the controversy is such that wondering why – not whether – most published findings are false [2] is an urgent question, for example in the biomedical research field, where harmful consequence to health may arise. This question is formulated in terms of so called false discoveries in the context of simultaneous testing of hypotheses in experimental research.

In this paper Fisher's and Neyman–Pearson's theories developed in classical statistics are focused from a logical point of view; related problems are presented and discussed, and the inference approach based on Bayesian statistics is introduced. The state of the art of multiple hypothesis tests is reviewed with special attention to the problem of estimation and control of false discovery rate (FDR) with applicability to bio-metrology.

2. Frequentist and Bayesian rationals of statistical inference

2.1. Classical approaches

In the so-called frequentist area of interpretation of probability, Fisher [3] advocated the use of *p*-values in significance testing of so-called null hypothesis (usually denoted by H_0) against the Neyman and Pearson (N–P) theory [4] that introduced the concept of an alternative hypothesis to the null one, and the probabilistic notions of type I/type II (false positive/negative) errors. Both Fisher's and N–P's methods disregarded the interplay of prior and conditional probabilities: this point was taken into account by von Mises in his approach [5] developed in terms of success or error chance of the test (von Mises' use of the term 'chance' rather than probability was intended to point out that no randomization was required in his approach) starting from the assumption of the need of a prior distribution function. Fisherian and N–P theories of testing hypotheses (and related theories of point or interval estimation) were set by Wald in the overall context of the problem of statistical inference [6].

2.2. Role of posterior probability

An early Bayesian approach to this problem was developed by Jeffreys [8]. The conflict between frequentist interpretation and Bayesian interpretation of statistical testing was modelled by Lindley as a paradox showing how the posterior probability of an hypothesis conditional on experimental result e, $p(e \mid H)$, is affected by prior probability $p(H)$: for quite small values of $p(H)$, $p(e \mid H)$ may become as high as 0.95, while a significance test may state that result e is significant for H at the $\alpha = 0.05$ level (in terms of p-value, values $p(e \mid H) \leq 0.05$ are compatible with $p(H \mid e) = 0.95$, being dependent on assigned values of prior $p(H)$ as per inverse probability calculation) according to Bayes' formula:

$$p(H \mid +) = p(H)p(+ \mid H)\left[p(H)p(+ \mid H) + p(\neg H)p(+ \mid \neg H)\right]^{-1} = 1 - p(\neg H \mid +) \qquad (1)$$

For example, the characteristics of two diagnostic tests are shown in table 1, where the hypothesis tested is $H \doteq$ 'disease is present' and the sign + (–, respectively) stands for 'test result is positive (negative, respectively)'. The significance level is α=0.05 for test1 (0.04 for test2) and the power is $1 - \beta = p(- \mid \neg H) = 0.98$ (for test2: 0.99). Power may be translated into test specificity: thus $p(- \mid \neg H)$ is an index of accuracy in making the negative diagnosis (i.e., correctly reporting absence of disease); $p(+ \mid H) = 1 - \alpha$ is an index of the accuracy in making the positive diagnosis, called the test sensitivity: with data in table 1, for test1 $p(+ \mid H) = 0.95$ (for test2: 0.96).

Table 1. Tests for the presence of a disease (H).

	Test results			
	Positive, +		Negative, –	
	test1	test2	test1	test2
H	0.95	0.96	0.05	0.04
$\neg H$	0.02	0.01	0.98	0.99

Table 2. Test2: numerical example, $p(H)$=0.005.

	+	–	total
H	48	2	50
$\neg H$	99	9851	9950
total	147	9853	10000

A rationale diagnostic criterion is to require $p(H \mid +) > p(\neg H \mid +)$. Both tests exhibit reliable performance. For test2, comparatively more sensitive and specific, baseline condition to obtain $p(H \mid +) > 1/2$ is $p(H) > 0.01$, i.e. a disease prevalence greater than 1%. Note that rare diseases do not satisfy such a condition (Poisson distribution is the probability paradigm of rarely occurring events). More generally, experiments designed to detect any rare events, for which baseline percentage is under the threshold of 1–2%, should take into account dependence of inverse probability on prior probability – other factors in Eq. (1) being equal. For example, using test2 with $p(H) = 0.005$, Eq. (1) yields $p(\neg H \mid +) \cong 0.675$: in this case, although test2 is quite sensitive (type I error probability is 0.04) and very specific (type II error probability is 0.01), the probability of an incorrect diagnose is almost twice the probability of a correct one. This is known as the false-positive paradox. Moreover, the probability of missed positives (false negatives) $p(H \mid -)$ results slightly larger than 0.0002. Table 2 reports an approximate experimental realization of test2, where the relative frequency of correct diagnosis is $48/147 \approx 0.33 \approx 1 - p(\neg H \mid +)$ and the relative frequency of false negatives is $2/9853 \approx 0.0002 \approx p(H \mid -)$. Data in table 2 show accuracy values in accordance with declared characteristics of test2: namely 0.96 of sensitivity and a specificity of approximately 0.99.

2.3. Controversy: from theory to practice

However this is not a paradox, as it follows from Bayes' formula; the conflict arises in acceptance of an inference criterion or another: is the significance of H best expressed by $p(e \mid H)$ or by $p(H \mid e)$?

According to Bayesian scholars, a strict logical line of reasoning leads to the probability of the hypothesis conditional on experimental result, i.e. $p(H|e)$ – the other one being flawed in principle. On the other hand, frequentist scholars insist on p-value $p(e|H)$; this position is stated in famous Fisher's statement: "The force with which such a conclusion is supported [by a significantly small $p(e|H_0)$] is logically that of a simple disjunction: *Either* an exceptionally rare chance has occurred [in our symbols: e] *or* the theory of random distribution [H_0] is not true" [3: p. 42]. (Crediting the first alternative, H_0 may show what can be miscalled its trueness, after the occurrence of an unlikely event.) Ever since, the controversy among schools of thought has been continuously fed from diverse theoretical perspectives (see e.g. [9–18]) and in view of a variety of applied statistics fields, including signal detection techniques (false vs. missed alarms), social sciences (e.g., criminal-court cases), therapeutic industry (e.g., new treatment efficacy vs. placebo), and metrology (outliers in measurements). A thorough review of uses and misuses of null hypothesis significance testing can be found in [19], proposed by "the editor of an empirical journal […] in attempting to develop a policy that would help ensure the journal did not publish egregious misuses" [19: pp. 241–242]. More recently, Fisher's, N–P's and Jeffrey's approaches were compared and contrasted in [20].

3. False discoveries in multiple hypothesis tests

3.1. Rise and consequence of the FDR problem

Problems with statistical significance become more complex when a multiplicity of hypotheses is under test. Tukey's account of the philosophy of multiple comparison is introduced by a very severe criticism [21: p. 100]: "Statisticians classically asked the wrong question and were willing to ask with a lie […]". Multiple hypotheses testing is of current interest to applied researchers involved in almost all of scientific disciplines [22]; in biology, e.g. with application to microarray experiments and bioinformatics [23–27]. An issue that gained attention in terms of false discovery can be described after the formulation by Sorić [28]. A provisional rejection of a null hypothesis is called a discovery, false discoveries are related to type I errors; let a large number n of independent experiments be contextually performed at a significance level α each, with an unknown large number N_0 of "true" nulls, and let $\rho=\alpha N_0 r$ denote the quotient of false discoveries among r declared ones: if almost all of those r discoveries are reported in scientific publications, it can be inferred that a non negligible – unless ρ is very small – part of research work may have been misled by false discoveries. Sorić in the above cited paper of 1989 [28] credits Edwards, Lindman, and Savage [29] for having drawn attention to the probability of false discoveries, although diversely alluded to in their Bayesian work published in 1963. However, according to Seeger [30], Eklund in three unpublished reports (in Swedish) archived in 1961–1963 at Uppsala University Institute of Statistics under the title "*Massignifikansproblemet*" already took into account and proposed solution to a problem relevant to large exploratory investigations: to keep low the proportion of predicted number of false significances to the observed number of significances – that was called by Eklund the "mass-significance" problem. However, pioneering work on this matter can be traced back towards the middle of the nineteenth century (references can be found to Cournot [31]: see e.g. [32]). A classical method – proposed by Bonferroni in 1935 [33] (as cited e.g. in [34]) – is known with the name of Bonferroni's correction.

3.2. FDR: estimation and control methods

The main problem in testing a family of hypotheses at a time, is that the overall significance level increases with the number of hypotheses. Suppose the significance level for a single hypothesis is α, and that n such hypotheses are under test: if these statistical hypotheses are mutually independent, the overall significance level is $1–(1– \alpha)^n$, that for small α is approximately $n\alpha$. Bonferroni's correction leads to downscale the individual α by a factor $1/n$ (thus n times the p-value computed for each hypothesis, the so called adjusted p-value, is checked not to exceed α), so that the overall significance level is closely reset to value α. Improving on proposed solutions, the problem to bound the overall type I error has been approached in diverse terms, mainly based on estimating and controlling its

probability, or the expected value of its relative frequency or its rate (see Shaffer [32] for a thorough review), till in 1995 Benjamini and Hochberg [35] introduced the formulation of a false discovery rate (FDR) and developed a methodics to control FDR in multiple testing. More historical notes are reported by Benjamini and Hochberg [36]. Another historical perspective is given by Tukey [21]. A unified (frequentist/Bayesian) approach to FDR estimation can be found in [34].

Table 3 summarizes results of multiple test of n hypotheses, of which N_0 are "true" null. R are declared discoveries. N_0, R, U: random variables (RVs), n_0, r, u respective realizations; n is known.

Table 3. U/N_0 is the "false positive" fraction, U/R the fraction of "false discoveries".

	Hypotheses supported by test		
	null	alternative	total
"true" null	$N_0–U$	U	N_0
"false" null	$n–R–(N_0–U)$	$R–U$	$n–N_0$
total	$n–R$	R	n

The FDR can be expressed in the following form, where the symbol $\vee$ in the denominator signifies that in case R=0 the expected value, E[·], is set to zero (to overcome the problem of the division 0/0).

$$\mathrm{FDR} = \mathrm{E}\left[U/\{R \vee 1\}\right] \tag{2}$$

Conditioning on R>0, a positive FDR is obtained [37]:

$$\mathrm{pFDR} = \mathrm{E}\left[\frac{U}{R} \middle| R > 0\right] \tag{3}$$

Diverse control methods (single- or multi-stage) have been proposed to guarantee a FDR under a preset overall significance level, say q. It should however be stressed that FDR control is involved with an expected rate, thus some combinations of observed values of U=u and R=r>0 may result into $u/r>q$, even if pFDR≤q. Benjamini and Hochberg [35, 36] introduced a method based on reordering the p-values observed in n simultaneous tests; to distinguish them from the p_i before reordering (when the subscript individuates the i-th test), indices of p-values after monotonic reordering are bracketed:

$$p_{(1)} \le \cdots p_{(j)} \cdots \le p_{(n)}, \quad 1 \le j \le n \tag{4}$$

This gives rise to a step-up procedure, for each $p_{(j)}$ is matched against the quantity jq/n; eventually, it is defined the index $J = \max\left(j : p_{(j)} \le jq/n\right)$ that must be estimated, say by $\hat{J}$. After that, all the hypotheses pointed to by the indices $j \le \hat{J}$ are rejected. Finally, the so-called Benjamini–Hochberg (BH) rule for the correction of the original p_i-values, results into the BH–corrected values, where index(p_i) denotes the position (j) of the observed p_i in the succession of Eq. (4):

$$p_i^{\mathrm{BH}} = np_i/\mathrm{index}\left(p_i\right) \tag{5}$$

A twofold insight into the possible usage of FDR can be articulated. On one hand, a threshold on can be pre-set and a multiple test be designed so that the attainable FDR level does not exceed the threshold: this approach is aimed at performing what is appropriately called FDR control [32, 35].

On the other hand, after having fixed the acceptance p-values threshold, a point-wise estimator of the FDR is constructed so that the expectation of this estimator is not less than the FDR value calculated at that threshold: this is the estimation approach. A point estimation approach can be summarized starting from Eq. (2). Putting $U(\alpha) = \#\{\text{"true" null hypotheses} : p_i \le \alpha\}$ and $R(\alpha) = \#\{\text{null hypotheses} : p_i \le \alpha\}$, where the operator $\#\{\cdot\}$ returns the cardinality of the set defined

inside brackets: $\mathrm{F_{DR}}(\alpha)=\mathrm{E}\left[U(\alpha)/\{R(\alpha)\vee 1\}\right]$. According to Storey [37], an estimator of $\mathrm{FDR}(\alpha)$ for a given α is $\hat{\mathrm{F}}_{\mathrm{DR}}(\alpha)=\alpha\hat{N}_0(\tau)/\{R(\alpha)\vee 1\}$, where τ is a tuning parameter and $\hat{N}_0(\tau)$ is the estimate of total number N_0 of "true" null hypotheses:

$$\hat{N}_0(\tau)=(n-R(\tau))/1-\tau \tag{6}$$

The estimator of Eq. (6) is justified by approximate equalities $\alpha\hat{N}_0(\tau)\cong U(\alpha)$, $R(\alpha)\vee 1\cong R(\alpha)$: it was shown [37] that conditions can be stated in order to grant that the inequality $\mathrm{E}\left[\hat{\mathrm{F}}_{\mathrm{DR}}(\alpha)\right]\geq \mathrm{F_{DR}}(\alpha)$ holds: based on this estimator $\hat{\mathrm{F}}_{\mathrm{DR}}(\alpha)$, a unified approach to FDR control and estimation is presented in [38]. More on unified approaches can be found in [34].

A Bayesian approach is illustrated in terms of a so-called two-class model. It assumes that a set of (null) hypotheses $\{H_{0\mathrm{i}}:1\leq i\leq n\}$ under test can be represented by identically distributed RVs – supposed sharing a Bernoullian distribution, say H_0– associated to the same test statistics ϑ with significance region ω. By assumption, the set can be bi-partitioned: each $H_{0\mathrm{i}}$ is assigned to one or the other partition according to its truth state – either "true" or "false". Let associate H_0 to the binary indicator $\ddot{H}_0$, such that $\ddot{H}_0=0$ with probability $\Pr(\ddot{H}_0=0)=\pi_0$ if H_0 is "true", otherwise $\ddot{H}_0=1$, $\Pr(\ddot{H}_0=1)=1-\pi_0$ (the frequency n_0/n can be an empirical estimate of prior π_0). Storey [39] proves a theorem, equating a "posterior Bayesian type I error probability function" to pFDR (Eq. (3)):

$$\Pr(\ddot{H}_0=0\mid\vartheta\in\omega)=\frac{\pi_0\Pr(\vartheta\in\omega\mid\ddot{H}_0=0)}{\pi_0\Pr(\vartheta\in\omega\mid\ddot{H}_0=0)+(1-\pi_0)\Pr(\vartheta\in\omega\mid\ddot{H}_0=1)}=\mathrm{E}\left[\frac{U(\vartheta)}{R(\vartheta)}\middle|R(\vartheta)>0\right] \tag{7}$$

4. Conclusion

The problem of hypothesis testing about the (unknown) magnitude of a quantity x being estimated (in measurement science term: the measurand) is the reverse of the coin of interval estimation of x along with its measurement uncertainty. However there is no (neither in Bayesian nor in frequentist statistical inference) *experimentum crucis* to discriminate "true" hypotheses from "false" ones. This is a complex problem that rises even extra-difficulties when a multiplicity of hypotheses are tested for significance at a time, as it happens in a variety of bio-scientific experiments.

Techniques for estimation and control of the probability of incurring in false discoveries (i.e., test results erroneously qualified as significant) are available to reduce to manageability – on the average – such a complexity. This paper was focused on main points selected from foundational issues and state of the art developments (a rich bibliography can be found in [40]).

Coverage probabilities of multiple confidence intervals will be the matter of next efforts addressed to evaluation and expression of uncertainties of a multiplicity of quantities under interval estimation.

References

[1] Brumfiel G 2008 Significant *Nature* **455** 1027
[2] Ioannidis J P A 2005 Why Most Published Research Findings Are False *PLoS Medicine* **2** 696
[3] Fisher R A 1973 *Statistical Methods and Scientific Inference* 3rd ed. (London: Macmillan)
[4] Neyman J and Pearson E S 1933 On the Problem of the Most Efficient Tests of Statistical Hypotheses *Philos. Trans. Roy. Soc. London Ser. A* **231** 289
[5] Mises R v 1943 On the Problem of Testing Hypotheses, *Ann. Math. Stat.* **14** 236
[6] Wald A 1942 *On the Principle of Statistical Inference* (Notre Dame, Ind.: Uni. of Notre Dame)
[7] Lindley D V 1957 A Statistical Paradox *Biometrika* **44** 187
[8] Jeffreys H 1961 *Theory of Probability* 3rd ed. (Oxford: Clarendon Press), chaps. 5–6
[9] Gumbel E J 1943 On the Reliability of the Classical Chi-Square Test *Ann. Math. Stat.* **14** 253
[10] Rosenkrantz R D 1973 The Significance Test Controversy *Synthese* **26** 304

[11] Birnbaum A 1977 The Neyman-Pearson Theory as Decision Theory and as Inference Theory; with a Criticism of the Lindley-Savage Argument for Bayesian Theory *Synthese* **36** 19
[12] Cox D R 1977 The Role of Significance Tests *Scand. J. Statist.* **4** 49
[13] Cohen Y 1994 The Earth Is Round ($p<.05$) *American Psychologist* **49** 997
[14] Royall R M 1997 *Statistical Evidence: A Likelihood Paradigm* (London: Chapman & Hall)
[15] Breaugh J A 2003 Effect Size Estimation: Factors to Consider and Mistakes to Avoid *J. of Management* **29** 79
[16] Nakagawa S and Cuthill I C 2007 Effect Size, Confidence Interval and Statistical Significance: A Practical Guide for Biologists *Biol. Rev*. **82** 591
[17] Li X R and Li X-B 2008 Common Fallacies in Applying Hypothesis Testing *Proc. 11th IEEE Conf. on Information Fusion* (Cologne, Germany: IEEE)
[18] D'Errico G E 2009 Testing for Outliers Based on Bayes Rule *Proc. IMEKO XIX World Congress–Fundamental and Applied Metrology* (Lisbon, Portugal: IMEKO) pp. 2368–2370
[19] Nickerson R S 2000 Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy *Psychological methods* **5** 241
[20] Berger J O 2003 Could Fisher, Jeffreys and Neyman Have Agreed on Testing? (with comments and rejoinder) *Statistical Science* **18** 1
[21] Tukey J W 1991 The Philosophy of Multiple Comparisons *Statistical Science* **8** 100
[22] Gelman A, Hill J and Yajima M 2012 Why We (Usually) Don't Have to Worry About Multiple Comparisons *J. Research on Educational Effectiveness* **5** 89
[23] Dudoit S, Shaffer J P and Boldrick J C 2003 Multiple Hypothesis Testing in Microarray Experiments *Statistical Science* **18** 71
[24] Storey J D and Tibshirani R 2003 Statistical Significance for Genomewide Studies *Proc. Natl. Acad. Sci. USA* **100** 9440
[25] Efron B 2007 Size, Power and False Discovery Rates *Ann. Statist.* **35** 1351
[26] Efron B 2008 Microarrays, Empirical Bayes and the Two-Groups Model *Ann. Statist.* **23** 1
[27] Langaas M, Lindqvist B H and Ferkingstad E 2005 Estimating the Proportion of True Null Hypotheses, with Application to DNA Microarray Data *J. R. Statist. Soc.* B **67** 555
[28] Sorić B 1989 Statistical 'Discoveries' and Effect-size Estimation *J. Am. Statist. Ass.* **84** 608
[29] Edwards W, Lindman H and Savage L J 1963 Bayesian Statistical Inference for Psychological Research *Psychological Review* **7** 193
[30] Seeger P 1968 A Note on a Method for the Analysis of Significances en Masse *Technometrics* **10** 586
[31] Cournot A A 1843 *Exposition de la Théorie des Chances et des Probabilités* (Paris: Hachette)
[32] Shaffer J P 1995 Multiple Hypothesis Testing *Annu. Rev. Psychol* **46** 561
[33] Bonferroni C E 1935 Il calcolo delle assicurazioni su gruppi di teste *Studi in Onore del Professore Salvatore Ortu Carboni* (Roma, I) pp. 13–60
[34] Strimmer K 2008 A Unified Approach to False Discovery Rate Estimation *BMC Bioinformatics* **9** (14 pages, online, doi:10.1186/1471-2105-9-303)
[35] Benjamini Y and Hochberg Y 1955 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing *J. R. Statist. Soc.* B **57** 289
[36] Benjamini Y and Hochberg Y 2000 On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics *J. Educational and Behavioral Statistics* **25** 60
[37] Storey J D 2002 A Direct Approach to False Discovery Rates *J. R. Statist. Soc.* B **64** 479
[38] Storey J D, Taylor J E and Siegmund D 2004 Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach *J. R. Statist. Soc.* B **66** 187
[39] Storey J D 2003 The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value *Ann. Statist.* **31** 2013
[40] Rao C V and Swarupchand U 2009 Multiple Comparison Procedures–a Note and a Bibliography *J. of Statistics* **16** 66