INSTITUTE OF PHYSICS PUBLISHING CLASSICAL AND QUANTUM GRAVITY

Class. Quantum Grav. **22** (2005) S1373–S1379 doi:10.1088/0264-9381/22/18/S50

False discovery rate: setting the probability of false claim of detection

L Baggio[1] **and G A Prodi**

Dipartimento di Fisica, Università di Trento and INFN, Gruppo collegato di Trento, Sezione di Padova, via Sommarive, 14, 38050 Povo, TN, Italy

E-mail: baggio@science.unitn.it

Received 12 April 2005, in final form 29 July 2005
Published 6 September 2005
Online at stacks.iop.org/CQG/22/S1373

Abstract

When testing multiple hypotheses in a survey—e.g. many different source locations, template waveforms, and so on—the final result consists of a set of confidence intervals, each one at a desired confidence level. But the probability that at least one of these intervals does not cover the true value increases with the number of trials. With a sufficiently large array of confidence intervals, one can be sure that at least one is missing the true value. In particular, the probability of false claim of detection becomes non-negligible. In order to compensate for this, one should increase the confidence level, at the price of reduced detection power. False discovery rate control (Benjamini Y and Hochberg Y 1995 *J. R. Stat. Soc.* B **57** 289–300) is a relatively new statistical procedure that bounds the number of mistakes made when performing multiple hypothesis tests. We shall review this method, discussing exercise applications to the field of gravitational wave surveys.

PACS numbers: 04.80.Nn, 02.50.Cw, 95.75.−z

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The motivation for controlling the false discovery rate (FDR)—i.e. the fraction of false alarms in a collection of candidate detections—came to our attention as we were involved in data analysis for the IGEC [3], the network of resonant detectors that searched for coincident burst gravitational wave (GW) signals in the years 1997–2000. Even if the detectors involved in IGEC were rather similar, there were obvious configurations (special choice of detector pairs, three-fold instead of double coincidence) or cuts of the data (higher or lower threshold

[1] Present address: Institute for Cosmic Ray Research, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken, 277-8582, Japan.

Table 1. Quick-reference notation chart for the variables used in section 2. m is the total number of performed tests (trial factor), m_0 and m_1 are the real number of underlying off-source and on-source tests. The number of *actually* positive tests is R, given by S true positives and B spurious claims. An ideal experiment would neither treat background as signal (type I error) nor do the reverse (type II errors).

	Null retained (cannot reject)	Reject null (i.e. accept alternative)	Total
Null (H_0) true (background)	$m_0 - B$	B Type I error	m_0
Alternative true (signal)	$\beta = m_1 - S$ Type II error	S Detected signals	m_1
	$m - R$	$R = B + S$ Reported signal candidates	m

on event amplitude) characterized by lower background counts, or higher duty time. We did not have *a priori* a good reason to prefer one configuration or cut more than others, as we did not know *a priori* the intensity of the signal, and hence the efficiency. Therefore, we decided at the beginning on a fairly long list of interesting choices, in order to perform many analyses in parallel and eventually to quote the results for each trial.

The results were expressed as confidence intervals on the expectation value for the number of counts in coincidence due to GW. When unveiling the final results, one of the confidence intervals at 90% coverage was not including the null hypothesis (i.e. zero counts). Of course, this can be somewhat expected by chance when the number of trials is very high. It was possible to compute accurately that with 30% probability there was a chance that at least one of the tests falsely rejected the null hypothesis.

The probability of at least one false claim in a set of trials is known as the *family-wise error rate* (FWER). It is not difficult to devise a method to control this quantity *before* going to the results: we just have to increase the confidence in the single trial (say 99%, or 99.99% coverage) in order to keep the FWER much lower than one. The drawback is that the resulting confidence interval would be much larger, and consequently the power of the search would fall dramatically. This is a consequence of the request that *not even in a single case* the null hypothesis is rejected when it is true.

A very reasonable compromise was suggested by Benjamini and Hochberg [1]. They remark that in many practical cases, when having one or more false claim is not by itself unacceptable, we could just be happy if—on average—*most* of the claims were real. In other words, they propose to bound FDR instead of FWER.

There are many topics in the GW search which would benefit from this kind of procedure, for instance:

- all sky surveys: many source directions and polarizations are tried in parallel;
- template banks;
- eyes-wide-open searches: many alternative analysis pipelines, with different amplitude thresholds, signal duration and so on, are applied on the same data;
- periodic updates of results: every new science run is a chance for a 'discovery' ('maybe the next one is the good one');
- Many graphical representations or aggregations of the data ('with a slight change in the binning, the 'signal' shows up better').

This work does not mean to be a complete review of the state-of-art techniques of FDR control, but hopefully it will be a stimulus for whoever is involved in multiple-test data analysis issues.

In the following sections, we shall use the notation reported in table 1.

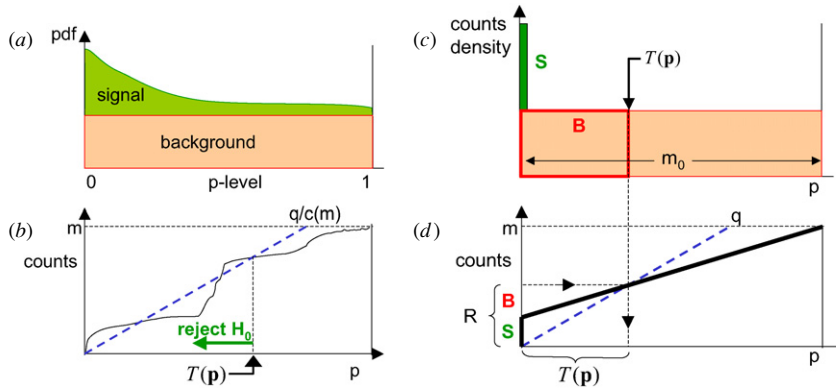

Figure 1. (*a*) The probability density function of *p*-values when data come from a mixed model can be thought of as the sum of a uniform distribution (background) and a biased one (signal). (*b*) The Benjamini–Hochberg procedure (BH) consists in plotting the cumulative histogram of the *p*-values of the m trials (continuous line) and looking for intersections with a line drawn from the origin and with slope equal to $m \cdot c(m)/q$ (dashed line). The null hypothesis is rejected for all data with *p*-value between 0 and the abscissa of the highest intersection point. (*c*) Sketch of histogram of *p*-values and (*d*) corresponding cumulative histogram, in a case of easily separable signals. The BH procedure applied to this case can easily be shown to control FDR (see section 2.3 for details).

2. Description of the method

2.1. Preliminary remarks

In order to decide whether the results of a measurement are compatible with being generated by noise only (*null hypothesis*, H0) or instead they contain a signal (*alternative hypothesis*, H1) the textbook procedure is to set up a test statistic t from the measures themselves. If $F_0(t)$ is the distribution of t when the H0 holds, then the *p*-value of t is defined as $p = F_0(t) = \Pr(t_0 > t | \forall\, t_0)$. By construction p is uniformly distributed between 0 and 1:

$$\Pr(p < p_0 | 0 \leqslant p_0 \leqslant 1) = p_0. \tag{1}$$

It is of paramount importance that the distribution F_0 is known. It is always wise to check *a priori* models with a goodness-of-fit test, when there are enough off-source data available. This is not always the case, but often there are surrogate procedures (e.g. data permutation) which give fresh independent samples of the background process, removing at the same time the effect of real signals, if any are present in the data. For instance, in the case of IGEC, the *resampling procedure* consisted in adding a delay to the time reference of one of the detectors in the network, such that the coincident signal is lost, while the background expectation value of coincidence counts is approximately unchanged. In the case the data are not compliant with the model, at worse resampled data may allow us to estimate F_0 by empirical fit.

As for H1, it is usually unknown, but for our purposes it is sufficient to assume that the signal can be distinguished from the noise, i.e. $\Pr(p < p_0 | 0 \leqslant p_0 \leqslant 1) \neq p_0$. The sketch in figure 1 (top left) illustrates the concept.

For a single hypothesis test, the condition 'reject null if $p < \alpha$' leads to false positives with probability α. In the case of multiple tests, we deal with a set $\mathbf{p} \equiv \{p_1, p_2, \ldots, p_m\}$ of *p*-levels, which need not to be derived from the same test statistics, nor they should refer to the same tested null hypothesis. m is called the trial factor. We select discoveries using a threshold $T(\mathbf{p})$: 'reject null if $p_j < T(\mathbf{p})$'.

2.2. Controlling type I errors (B)

The *uncorrected testing* would just use the same threshold for each test: $T(\mathbf{p}) = \alpha$. The probability that at least one rejection is wrong grows as $P(B > 0) = 1 - (1 - \alpha)^m \approx m\alpha$.

Therefore, as in the IGEC case, false discovery is guaranteed for large enough m.

The other extreme solution, usually referred to as the *Bonferroni procedure* [4], controls the FWER in the most stringent manner, by requiring that $P(B > 0) \leqslant \alpha$. This is achieved by the choice $T(\mathbf{p}) = \alpha/m$ While this approach makes mistakes rarely, the cost is low efficiency ($S \approx 0$).

2.3. Controlling the false discovery rate (B/R)

In order to trade between $B = 0$ and $S = 0$, the FDR control focuses on the ratio of false discoveries to the total number of claims:

$$\mathrm{FDR} \equiv \begin{cases} B/R \equiv B/(B+S) & \text{if} \quad R > 0 \\ 0 & \text{if} \quad R = 0. \end{cases} \tag{2}$$

This can be done with a proper choice of $T(\mathbf{p})$. The original procedure suggested by Benjamini and Hochberg (BH) is extremely simple, involving only trivial algebraic operations. It consists of the following steps.

- sort the p-values in ascending order: $\{p_1, p_2, \ldots, p_m | i < j \Rightarrow p_i \leqslant p_j\}$;
- choose your desired FDR q (in case no signal source is actually present during the observation, then the procedure is equivalent to the Bonferroni procedure with $\alpha = q$);
- define $c(m) = 1$ if p-values are independent or positively correlated; otherwise $c(m) = \sum_{j=1}^{m} 1/j$;
- determine the threshold $T(\mathbf{p}) = p_j$ by finding the index j such that $p_k > k(q/m)/c(m)$ when $k > j$ (see figure 1 for a visual representation of this condition).

The above procedure with $c(m) = 1$ was shown [1] to control the expectation value[2] of the FDR at least at level q in the case when all m tests are independent. However, it was later proved to control the FDR when tests are positively correlated [2] (for instance, multivariate normal data where the covariance matrix has all positive elements). The alternative definition of $c(m)$ given above controls the FDR in the most general case [2], but at the cost of reduced efficiency.

There is a nice back-of-the-envelope plausibility argument which can be found in [5] for the simple case when signals are easily separable (e.g. signals with high signal-to-noise ratios). In this case we expect their p-level to be very low, and correspondingly in the cumulative histogram of p-levels we shall see a step with height S near $p \approx 0$; see figure 1 (bottom right). We also see that there is only one intersection point for the BH procedure, such that

$$T(\mathbf{p})/R = q/m. \tag{3}$$

On the other hand, the threshold $T(\mathbf{p})$ can be expressed on average by B/m_0 (this is a special case of (1)). Substituting this value into (3) we obtain

$$B/R = qm_0/m \leqslant q. \tag{4}$$

For a rigorous proof, see [2].

[2] Of course, the quantity FDR is a random variable, as well as the p-values.

Table 2. Results of the simulation described in section 3. The first column lists the values of N_s, the other columns refer to different values of N_b, as listed in the first row. Each entry corresponding to a $\{N_s, N_b\}$ couple is composed by values: the upper one refers to the Bonferroni procedure, the lower to the BH procedure. These values are averaged over 40 000 samples and the statistical precision is of the order of 0.005.

N_s	0.01	0.02	0.05	0.1	0.2	0.5	1	5	10	50
0	0.005	–	–	10^{-4}	3×10^{-4}	0.003	0.007	0.008	0.003	0.004
	0.005	2×10^{-4}	–	10^{-4}	3×10^{-4}	0.003	0.007	0.008	0.003	0.004
1	1.005	4×10^{-4}	0.001	0.002	0.005	0.013	0.028	0.012	0.004	0.005
	1.010	0.019	0.001	0.002	0.005	0.013	0.028	0.012	0.004	0.005
2	2.004	5×10^{-4}	0.002	0.004	0.009	0.021	0.047	0.016	0.005	0.005
	2.010	2.019	0.002	0.004	0.009	0.021	0.047	0.016	0.005	0.005
3	3.005	0.001	0.003	0.006	0.013	0.032	0.069	0.021	0.006	0.005
	3.010	3.019	0.004	0.006	0.013	0.032	0.069	0.021	0.006	0.005
4	4.005	0.001	0.004	0.008	0.017	0.043	0.086	0.027	0.007	0.006
	4.010	4.018	0.125	0.008	0.017	0.043	0.086	0.027	0.007	0.006
5	5.004	0.002	0.005	0.009	0.020	0.053	0.106	0.029	0.008	0.006
	5.009	5.018	5.046	0.009	0.020	0.053	0.106	0.029	0.008	0.006
6	6.004	0.002	0.006	0.013	0.024	0.061	0.124	0.034	0.010	0.007
	6.009	6.017	6.043	0.013	0.024	0.061	0.124	0.034	0.010	0.008

3. Numerical test of the method

We now demonstrate this procedure with a simple example. Suppose we are given the results of 50 counting experiments, labelled by the index i. Their background is modelled as a Poisson random variable, with the same[3] known expectation value N_b for all i.

We consider two possible cases: in the first one, we draw 50 independent measures, in the other case we generate correlation by summing neighbour bins (i.e., if n_c^i represent independent counts in the ith bin, then the 50 correlated counts $n_c'^i$ are defined as $n_c'^i = n_c^i + n_c^{i-1}$, where $n_c^0 \equiv n_c^{50}$). We investigated different background levels (from $N_b = 0.01$ to $N_b = 50$) and different number of detected signals ($N_s = 0 - 6$), assuming—for the sake of simplicity—that each bin can have either one or zero count due to true signals.

In order to decide the presence of a signal we use the one-tail Poisson probability for the expected number of counts in each bin. In tables 2 and 3 the results of a Monte Carlo simulation are shown for independent and correlated measures respectively. For each configuration (differing by average background and extent of true signals) we compute the average number of claims R, i.e. the number of bins for which the null hypothesis is rejected. We present the results for Bonferroni and BH tests, both tuned to bound the FWER at 1% when no signal is present.

Both procedures are working as expected, controlling the FWER and the FDR respectively at the desired level. For high background values they give, as expected, similar results. On the other hand, the efficiency of the Bonferroni procedure falls to zero for $N_b > 0.01$, while the BH procedure is still effective, up to $N_b = 0.05$ in this example.

In figure 2 we can visualize how the BH procedure manages to grasp the signals promptly, as the background level lowers (see also figure 1).

[3] To avoid degeneracy due to the discreteness of the test statistic (many results collapsing at the same p-values), we actually spread the background of the experiments in a range $\pm 1\%$ around N_b.

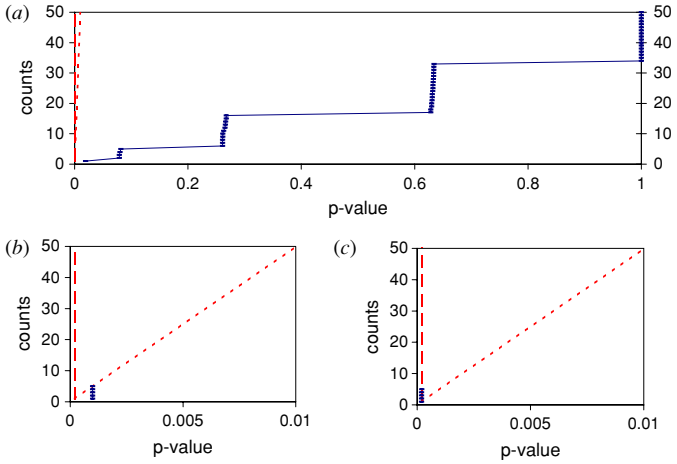

Figure 2. A few samples from the Monte Carlo used to produce table 2 are displayed in detail. They refer to $N_s = 5$, and the background is (*a*) $N_b = 50$, (*b*) $N_b = 0.5$, and (*c*) $N_b = 0.01$. In the above plots the cumulative histogram of the *p*-values is compared with the threshold given by the Bonferroni (– – –) and the BH (- - - -) procedures.

Table 3. Same as table 2 but for the case of correlated measures.

N_s	0.01	0.02	0.05	0.1	0.2	0.5	1	5	10	50
0	0.006	–	–	10^{-4}	2×10^{-4}	0.003	0.008	0.008	0.003	0.004
	0.010	0.010	–	10^{-4}	2×10^{-4}	0.003	0.009	0.008	0.003	0.004
1	1.005	3×10^{-4}	0.001	0.002	0.005	0.013	0.030	0.012	0.004	0.005
	1.010	0.029	0.001	0.002	0.005	0.013	0.031	0.012	0.004	0.005
2	2.005	7×10^{-4}	0.002	0.004	0.008	0.023	0.046	0.017	0.005	0.005
	2.009	2.018	0.006	0.004	0.008	0.023	0.047	0.017	0.005	0.006
3	3.004	0.001	0.003	0.005	0.012	0.032	0.067	0.022	0.006	0.005
	3.009	3.019	0.060	0.005	0.012	0.032	0.068	0.022	0.006	0.006
4	4.005	0.002	0.004	0.008	0.017	0.043	0.084	0.025	0.007	0.005
	4.009	4.017	0.143	0.008	0.017	0.043	0.085	0.025	0.007	0.006
5	5.004	0.002	0.005	0.010	0.019	0.051	0.107	0.029	0.008	0.007
	5.009	5.018	5.047	0.010	0.019	0.051	0.108	0.029	0.008	0.007
6	6.004	0.003	0.007	0.011	0.024	0.061	0.127	0.035	0.010	0.007
	6.009	6.016	6.044	0.013	0.024	0.061	0.127	0.035	0.010	0.007

4. Conclusions

When multiple tests are tried for the same data set, controlling FDR seems in general a wiser idea than just limiting type-I errors. Robust but simple procedures exist which (conservatively) control the FDR in positively correlated tests, and also in the more general case (but at the cost of reduced efficiency).

This idea is relatively new in the astrophysics community, and we are not aware of any application in the GW community. Its application should be encouraged. Note, however, that

the BH procedure is not the only one, and more complex—but approximate—strategies have been investigated (see, for instance, [7, 6]).

Acknowledgments

We are indebted to James T Linnemann (MSU) for introducing us to the FDR. LB acknowledges the hospitality of the ICRR and a grant from Tokyo University.

References

[1] Benjamini Y and Hochberg Y 1995 *J. R. Stat. Soc.* B **57** 289–300
[2] Benjamini Y and Yekutieli D 2001 *Ann. Stat.* **29** 1165–88
[3] Astone P *et al* 2003 *Phys. Rev.* D **68** 022001
[4] Hochberg Y and Tamhane A C 1987 *Multiple Comparison Procedures* (New York: Wiley)
[5] Miller C J *et al* 2001 *Astron. J.* **122** 3492–505
[6] Yekutieli D and Benjamini Y 1999 *J. Stat. Plan. Inference* **82** 171–96
[7] Storey J D 2002 *J. R. Stat. Soc.* B **64** 479–98