MULTIPLE COMPARISONS AND THE FALSE DISCOVERY RATE

ŚAUNAK SEN[1]

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

sen@biostat.ucsf.edu

INTRODUCTION

In today's world of high throughput technologies, making multiple comparisons is a necessity. Although statistical methods for dealing with multiple comparisons issues have been around for a while, they received a special impetus from the rise of microarray technologies. In this note we will review some methods for thinking about and dealing with multiple comparisons. We will begin with considering situations when we have and when we don't have to adjust for multiple comparisons. Then we will learn about two types of adjustments in the evidence in p-values – the Family Wise Error Rate (FWER), and the False Discovery Rate (FDR).

To adjust or not   Adjustments for multiple comparisons are needed to avoid spurious associations. If we make a large number of tests, then it is quite likely that we will find a small p-value just by chance. When we are making many tests, each for a *different* hypothesis, then no adjustments for multiple comparisons are needed. For example, we may be interested in two SNPs each in two separate candidate genes, on two separate pathways to a phenotype. In this case we do not need to adjust for multiple comparisons. However, if multiple tests are being made to test a common hypothesis, then some adjustment needs to be made. For example, in genome scans, we are test each possible location on the genome for linkage with the phenotype to establish if genetic variation in the family is associated with phenotypic variation. In this case, we need to adjust for multiple comparisons. Some scenarios may not be clear-cut. In these cases we may have to make cautious judgments.

There are broadly two types of multiple comparisons issues. Sometimes we are making a number of tests that are related to each other, for example, we may be comparing the effects of different levels of a treatment. A treatment may have four doses, and we may be making pairwise comparisons between doses. A different scenario is when we are making separate comparisons such as when we are testing different SNPs for association with a phenotype, or when we are testing if different gene expression levels (as measured on a microarray) are different in cases versus controls. We will consider the latter case in detail in this note.

Three examples   Let us consider three examples where multiple comparison adjustments have to be made.

Deviant search from mutagenesis   Mutagenesis scans attempt to detect unusual mice, which have been exposed to mutagens, using a battery of physiological tests. Each mouse is followed in cages to measure physiological and behavioral characteristics. We get p-values for each measurement for each mouse. The question is what p-value cutoff should be used to flag a deviant mouse?

Genome scans   In genome-wide association scans a large number of markers are scanned for association between the phenotype and the marker. For each we get a p-value. There may be correlations between neighboring markers. We have to find which markers are truly associated with the phenotype.

---

[1]Lecture notes for Molecular and Genetic Epidemiology II, April 26, 2005.

Table 1: Table of declared true/false, and actual hypotheses for a p-value cutoff t

		Declared		
		Null	Alt	Total
Actual	Null	U_t	V_t	m_0
	Alt	T_t	S_t	m_1
	Total	W_t	R_t	m

Differential gene expression   Microarrays measure genome-wide gene expression for thousands of genes in different experimental conditions or tissues. We have thousands of p-values , one for each gene, testing the null hypothesis of no association between transcript levels and experimental conditions. We have to decide, based on these p-values, which genes may be considered differentially expressed under the conditions considered.

FAMILY WISE ERROR RATE (FWER)

Assume that we have m independent tests to test, $H_1, H_2, \ldots, H_m$ to test. In each case the null hypothesis is either true (H_i=0) or not (H_1=1). We have a test statistic that has a continuous distribution under the null hypothesis, and gives us p-values for each of the hypotheses. Call these $p_1, p_2, \ldots, p_m$. evidence in these m tests. Our goal is to combine the evidence in these p-values without making too many mistakes. Table 1 tabulates the possibilities for a cutoff t.

   The total number of hypotheses tested is m of which the null hypothesis is true m_0 times and the alternative is true m_1 times. Let us suppose that we use a certain method to *declare* hypotheses to be null or alternative. Of the m_0 null hypotheses we correctly declare them to be null U times, and V times we make a Type I error of declaring the alternative to be true when it isn't. Of the m_1 alternative hypotheses, we correctly identify S of them, and T times we make a Type II error of incorrectly accepting the null hypothesis. The total number of hypotheses *declared* to be false is R. The rest W of them are *declared* to be null. Note that U, V, S, T, R, and W will vary from sample to sample and are considered *random variables.*

   One criterion for this is the Family Wise Error Rate (FWER), which is the probability of rejecting one or more of the hypotheses erroneously. This criterion is appropriate when we are using the multiple tests to test a single bigger hypothesis. The FWER is $P(V \geq 1)$. The power of this procedure is the average of S/m_1 which we denote by $E(S/m_1)$.

Bonferroni procedure   If we want the FWER to be α or smaller, then we can use the Bonferroni procedure which says that we should reject hypotheses whose p-values are α/m or smaller. The Bonferroni procedure has the advantage of being simple, and valid even when the tests we are considering are dependent. Equivalently, we reject a hypothesis if the p-value multiplied by m is less than α. This is sometimes referred to as the Bonferroni correction.

   For example, if we perform three tests and get p-values 0.1, 0.04, and 0.01, and we want to control FWER to be 5% or less, then we will reject the null hypothesis only if the smallest p-value is 0.05/3 or less. In this case, one p-value (0.01) is smaller, and therefore we can reject the null hypothesis controlling the FWER at 5%. The Bonferroni-corrected p-value of the *family* of tests is $0.01 \times 3 = 0.03$.

Fisher combination procedure   If the tests we are considering are independent, then we can improve upon the Bonferroni procedure. One method is the Fisher combination procedure. Suppose $p_1, p_2, \ldots, p_m$ are m p-values obtained from independent tests. Let $T = -2 \sum \log(p_i)$. This is called the combination statistic. We compare this to a χ^2 distribution with $2m$ degrees of freedom to get the combination p-value for the family of tests. In the example considered above the combination statistic is $T = 2(2.30 + 3.22 + 4.61) = 20.26$ which gives a p-value of 0.0025 when compared to a χ^2 distribution with 6 degrees of freedom.

FALSE DISCOVERY RATE (FDR)

The *concept* of the false discovery rate grew out of the realization that in some scenarios such as microarray experiments, one is interested in controlling the number of false positives compared to the total number of positives. This is the false discovery rate. In Table 1 it would be V_t/R_t (assuming that $R_t > 0$). We will now use some graphs to understand the process.

Figure 1 shows the distribution of the p-values under the null hypothesis, alternative hypothesis, and in practice when about half of the hypotheses are null, and the other half are not. If all hypotheses we test are null, then the p-values follow a uniform distribution and so the histogram will be flat. If they come from the alternative hypothesis, then we should see a preponderance of small p-values, which is reflected in the idealized triangle shaped histogram. The rightmost panel in Figure 1 is the observed histogram of p-values we can expect to see. We have to decide a cutoff based on this histogram which hypotheses to reject and which to accept. Figure 2 shows how the False Discovery Rate changes as we change the cutoff. In our idealized scenario, the false discovery rate increases as we increase our cutoff because a greater proportion of p-values will be coming from the null.

Two approaches   There are two main approaches. The first, due to Benjamini and Hochberg seeks to find a cutoff, t, given a target proportion α, such that

$$E\left(\frac{V_t}{R_t}, R_t > 0\right) \leq \alpha.$$

The second, due to John Storey, and closely related to that of Efron and Tibshirani, seeks to find the *q-value*, for a fixed cutoff t such that

$$q_t = P\left(\frac{V_t}{R_t} | R_t > 0\right).$$

Thus, the first approach estimates a cutoff given a target *fdr*; the second estimates the *fdr* given a cutoff.

Benjamini-Hochberg procedure   Let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ be the ordered p-values. If π_0 is the proportion of null hypotheses, then the cutoff t such that

$$E\left(\frac{V_t}{R_t}, R_t > 0\right) \leq \alpha$$

is

$$t = \max\left\{p_{(i)} : p_{(i)} \leq \left(\frac{i}{m}\right)\left(\frac{\alpha}{\pi_0}\right)\right\}.$$

This is graphically represented in Figure 3. In practice, we do not know π_0, so a conservative choice is $\pi_0 = 1$. The advantage of this procedure is that there is a simple algorithm which also controls FWER if all the hypotheses tested are null.[2] The procedure offers more power than Bonferroni correction and also holds under weak dependence of the tests. One should also note that the procedure controls the proportion of false positives among rejected hypotheses *averaged over all studies*.

Efron-Tibshirani-Storey approach   If F_0 and F_1 are the distribution functions of the p-values under the null and alternative hypothesis respectively, then by Bayes theorem the q-value, or FDR corresponding to a cutoff of t is

$$q(t) = \pi_0 F_0(t)/F(t),$$

where $F = \pi_0 F_0 + \pi_1 F_1$. We can estimate $F(t)$ from the empirical distribution of p-values, and $F_0(t) = t$ since F_0 is uniform. Thus

$$\widehat{q(t)} = \pi_0 t/\widehat{F(t)} \leq t/\widehat{F(t)}.$$

Sharper estimates possible by estimating π_0. This is shown graphically in Figure 4.

---

[2]If all the tests are from null hypotheses, that is, if $m = m_0$, then the FDR reduces to the FWER.

Deviant search example   The objective is to screen a large number of potentially mutant mice for pheno-typic characteristics of interest, many of them subtle and invisible. A large number of mice pass through the scan and a large number of possibly related tests are performed on them. Mice flagged from the screen are followed up. Investigators need to know which mice are most promising and why.

In a pilot experiment, Kevin Seburn measured three physiological variables by following mice for three days in a cage.

1. RER (Respiratory Exchange Ratio)
2. Vertical movement
3. Ambulatory movement

Each variable was aggregated in the light and dark periods. A total of six measurements per mouse were obtained. Three sets of mice were formed – a *control training* set, a *control test* set and a *mutant test* set with deviants known from other experiments. Table 2 shows the results from the experiment. We used the Benjamini and Hochberg method to control FDR at 10%. We can see that the false positive rates and power are at desirable rates. Applying the FDR works better than applying a Bonferroni correction.

Table 2: Results from mutant search experiment. The columns labeled "Mice" refer to when the tests were applied on the mice, where as the columns labeled "Tests" refer to when the tests were applied to each individual physiological measurement. We applied the Benjamini-Hochberg procedure targeting a 10% FDR. The number of mice (and tests) flagged as mutant by FDR is shown in the "FDR" column. The estimated cutoff is shown in the "p-value cutoff" column. The number of mice (and tests) flagged by a Bonferroni correction is shown in the "Bonferroni" column.

	Number		FDR		P-value	Bonferroni	
	Mice	Tests	Mice	Tests	Cutoff	Mice	Tests
Control training	24	144	0	0	0	0	0
Control Test	24	144	2	6	0.0016	2	6
Mutant Test	22	132	17	49	0.0371	12	28

Figure 1: Distribution of p-values under null, alternative, and mixture settings. The left panel shows the distribution of p-values under the null hypothesis when it is just a uniform distribution. The middle panel shows the distribution under an alternative, when it is peaked towards smaller p-values. The right panel shows the distribution under a mixture of the two scenarios, when about half the p-values come from testing the null, and the other half from testing the alternative.
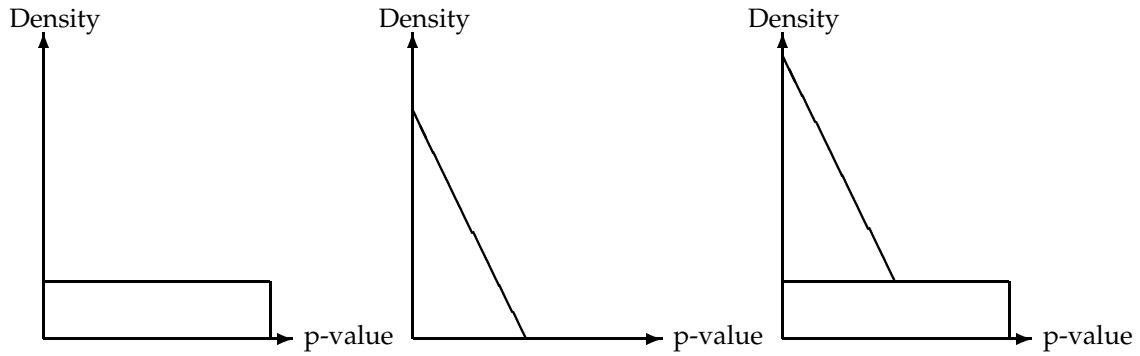
Density

p-value

Density

p-value

Density

p-value

Figure 2: How the false discovery rate varies with the cutoff.

Cutoff

Density

p-value

False positives
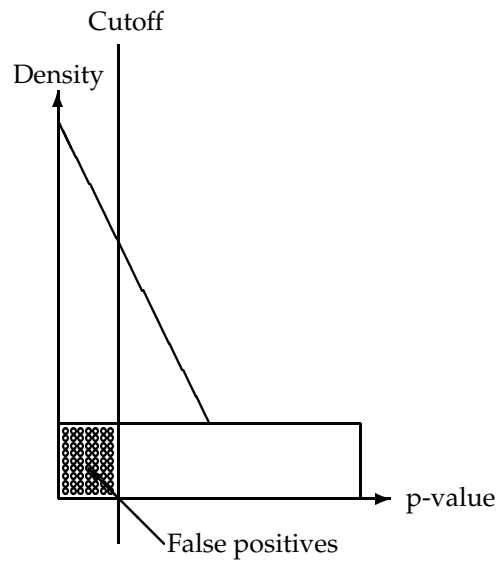
Figure 3: Graphic of Benjamini-Hochberg procedure.
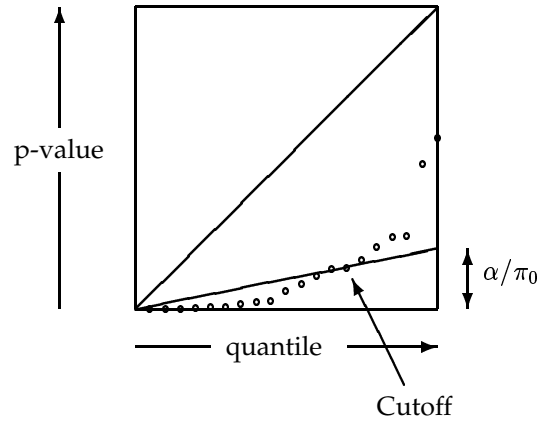
p-value

quantile

α/π_0

Cutoff

Figure 4: Graphic of Efron-Tibshirani-Storey procedure

Cutoff

Density

$F(t)$

$F_0(t)$

$\pi_0 F_0(t)$

p-value

False positives

6