Stat 215b (Spring 2004): Multiple testing and FDR

B. M. Bolstad
bolstad@stat.berkeley.edu

Feb 26, 2004

Multiple testing

Suppose we are carrying out multiple tests.

H_1 vs A_1 with p-value p_1

H_2 vs A_2 with p-value p_2

$\vdots$

H_m vs A_m with p-value p_m

If we knew which null hypotheses were true and which were not and we had some method of accepting or rejecting each tests (eg p-value < *alpha* our results would be something like the following table (of course in real life we never know whether the null is true or not).

	Not significant	Significant	Total
Null is True	U	V	m_0
Null is False	T	S	$m - m_0$
	$m - R$	R	m

Note that V is the number of type I errors and T is the number of type II errors.

If we were just doing traditional testing and looking at each hypothesis in isolation then we would just reject the null hypothesis H_i if $p_i < \alpha$. In isolation, the chance of making a type I error is just α. Unfortunately when you take into account that you are carrying out multiple tests the chance of making at least one type I error in m tests is $1 - (1 - \alpha)^m$. So if we carry out many hypothesis tests we are likely to get a sizeable number of false positives just by chance.

To account for this problem, the traditional approach is to use a Bonferroni adjustment to the significance level. In particular, we reject H_i if $p_i < \alpha/m$. However if m is large then α/m will become very small. This makes it hard to reject very many null hypotheses at all. Instead of having many false positives you will now have many false negatives.

FDR

Benjamini and Hochberg (1995) discuss a quantity known as the False Discovery Rate (FDR). In particular the FDR is

$$E\left(\frac{V}{R}|R > 0\right) P\left(R > 0\right)$$

They propose a procedure to control this quantity at a fixed level $q*$. Specifically first they order the P-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. Then they find the largest k such that

$$p_{(k)} \leq \frac{i}{m} q^*$$

then reject all $H_{(i)}$ for $i = 1, \ldots, k$.

Note that this is not the only method of controlling FDR (or related quantities). See for example the paper by Storey (2002).

What does all this mean? A simulation

Consider a microarray with 10,000 "genes". And suppose we have 5 arrays. For each gene on each array we have a log fold-change value. We will suppose that all the non-differential "genes" can be simulated as random normal variables with mean 0 and variance 1. The differential "genes" will be simulated as random normal variables with mean 5 and variance 1. This is a pretty extreme example. We will simulate 9000 non-differential and 1000 differential genes. We will carry out a $t - test$ of $H_0 : \mu = 0$ vs $H_0 : \mu \neq 0$ and compute the corresponding p-value for each gene. We will test at the 5% level of significance or control FDR at 5%. Then we will look at how many false positives and false negatives we get using each of the three methods.

The file `FDR.R` on the webpage carries out this simulation. For my run I computed 506 false positives and 0 false negatives using individual p values. Using the Bonferroni correction i had 0 false positives but also 812 false negatives. Controlling the FDR at 5% I had 80 False positives and 0 False Negatives.

From this simulation we see that using the FDR method reaches a balance somewhere between too many false positives and too many false negatives. You should run the simulation code with various settings for the difference in means and sample sizes to further explore the methods.