Multiple Comparisons:
Bonferroni Corrections and False Discovery Rates

Lecture Notes for EEB 581, ©Bruce Walsh 2004, version 14 May 2004

Statistical analysis of a data set typically involves testing not just a single hypothesis, but rather many (often *very* many!). For any particular test, we may assign a pre-set probability α of a type-1 error (i.e., a false positive, rejecting the null hypothesis when in fact it is true). The problem is that using a (say) value of $\alpha = 0.05$ means that roughly one out of every twenty such tests will show a false positive (rejecting the null hypothesis when in fact it is true). Thus, if our experiment involves performing 100 tests, we expect 5 to be declared as significant if we use a value of $\alpha = 0.05$ for each. This is the problem of multiple comparisons, in that we would like to control the false positive rate not just for any single test but also for the entire *collection* (or *family*) of tests that makes up our experiment.

How Many False Positives?

Suppose we perform n independent tests, each with a pre-set type one error of α. In this case, the number of false positives follows from the Binomial distribution, with α the probability of a "success" (a false positive) and n the number of trails. Hence, the probability of k such false positives is

$$\Pr(k \text{ false positives}) = \frac{n!}{(n-k)!\,k!}(1-\alpha)^{n-k}\,\alpha^k \tag{1}$$

For n large and α small, this is closely approximated by the Poisson, with Poisson parameter $n\alpha$ (the expected number of false positives),

$$\Pr(k \text{ false positives}) \simeq \frac{(n\alpha)^k e^{-n\alpha}}{k!} \tag{2}$$

Example 1. Suppose 250 independent tests are performed, and we have chosen a false-positive probability of $\alpha = 0.025$ (2.5%) for each. Suppose we observe 12 significant tests by this criteria. Is this number greater than expected by chance? Here, $n\alpha = 250\cdot 0.025 = 6.25$. The probability of observing 12 (or more) significant tests is

$$\sum_{k=12}^{250} \Pr(k \text{ false positives}) = \sum_{k=12}^{250} \frac{250!}{(250-k)!\,k!}(1-0.025)^{250-k}\,0.025^k$$

We could either sum this series (say in a spreadsheet) or simply recall the cumulative distribution function for a binomial, which is in **R**. In particular, the probability that a binomial with parameters n and p has a value of j or less is obtained by **pbinom(j,n,p)**. Hence the probability of 12 or greater is just one minus the probability of 11 or less, or **> 1- pbinom(11,250,0.025)**. **R** returns **0.02470269**. Given that there is only a 2.5% of this happening by chance, we expect some of these significant tests to be truly significant indeed, not false positives. The critical question, of course, is which ones?

Fisher's Method of Combining p Values Over Independent Tests

A interesting example of multiple comparisons is when the same hypothesis (i.e., smoking causes cancer) is independently tested. If we have the raw data, we can often combine all these experiments into a single data set. However, often this is not possible, either because the data are not fully reported, or the experiments are such that different variables are being followed and hence the raw data cannot be easily combined.

Fisher (1954) offered a simple, yet powerful, way around this based on the p values for each of the independent tests. If k independent tests (usually different studies from different groups) are performed, and the p value for test i is p_i, then the sum

$$-2\sum_{i=1}^{k}\ln(p_i) \tag{3}$$

approximately follows a χ^2_{2k} distribution. Fisher's method started the field of **meta-analysis**, wherein one searches the literature to find a number of tests of a particular hypothesis, and then combines these results into a single test. As Example 2 shows, none of the individual tests may be significant, but Fisher's combined approach can potentially offer more power, and hence can generate a significant result when all the tests are jointly considered. An import caveat to keep in mind during a literature search is the bias of reporting p values that are close to significant and not reporting p values that are far from significant.

Example 2. Suppose five different groups collected data to test the same hypothesis, and these groups (perhaps using different methods of analysis) report p values of 0.10, 0.06, 0.15, 0.08, and 0.07. Notice that none of these individual tests are significant, but the trend is clearly that all are "close" to being significant. Fisher's statistic gives a value of

$$-2\sum_{i=1}^{k}\ln(p_i) = 24.3921, \qquad \Pr(\chi^2_{10} \geq 24.39) = 0.0066$$

Hence, taken together these five tests show a highly significant p value.

Bonferroni Corrections and Their Extensions

Bonferroni corrections (and their relatives) are the standard approach for controlling the experiment-wide false positive value (π) by specifying what α values should be used for each individual test (i.e., we declare a test to be significant if $p \leq \alpha$). The probability of not making any type I (false positive) errors in n independent tests, each of level α, is $(1-\alpha)^n$. Hence, the probability of at least one false positive is just one minus this,

$$\pi = 1 - (1-\alpha)^n \tag{4a}$$

If we wish an **experiment-wide** false positive rate of π (i.e., the probability of one, or more, false positives over the entire set of tests is π), solving for the α value required for each test is

$$\alpha = 1 - (1-\pi)^{1/n} \tag{4b}$$

This is often called the **Dunn-Ŝidák method**. Noting that $(1-\alpha)^n \simeq 1 - n\alpha$, we obtain the **Bonferroni method**, taking

$$\alpha = \pi/n \tag{4c}$$

Both Equations 4b and 4c are often referred to as Bonferroni corrections. In the literature, π is occasionally referred to as the **family-wide error rate** (**FWER**), while α is denoted as the **comparison-wise error rate**, or **CWER**.

Example 3. Suppose we have $n = 100$ independent tests and wish an overall π value of 0.05. What α value to control for false positives should be used for each individual test? The Dunn-Ŝidák correction gives

$$\alpha = 1 - (1-0.05)^{1/100} = 0.000512$$

while the Bonferroni correction is

$$\alpha = 0.05/100 = 0.0005$$

Note that using such small α values greatly reduces the power for any single test. For example, under a normal distribution the 95% (two-side) confidence interval for the true mean is $\overline{x} \pm 1.96\sqrt{\text{Var}}$, while moving to an α value of 0.0005, $\overline{x} \pm 3.48\sqrt{\text{Var}}$.

Sequential Bonferroni Corrections

Under a strict Bonferroni correction, only hypotheses with associated p values $\leq \pi/n$ are rejected, all others are accepted. This results in a considerable reduction in power if two or more of the hypotheses are actually false. When we reject a hypothesis, there remain one fewer tests, and the multiple comparison correction should take this into account, resulting in so-called sequential Bonferroni corrections. Such sequential corrections have increased power, as Example 4 below shows.

Holm's Method

The simplest of these corrections is **Holm's method** (Holm 1979). Order the p values for the n hypotheses being tested from smallest to largest, $p(1) \leq p(2) \leq \cdots \leq p(n)$, and let $H(i)$ be the hypothesis associated with the p value $p(i)$. One proceeds with Holm's method as follows:

(i) If $p(1) > \pi/n$, accept all the n hypothesis (i.e., none are significant).

(ii) If $p(1) \leq \pi/n$, reject $H(1)$ [i.e., $H(1)$ is declared significant], and consider $H(2)$

(iii) If $p(2) > \pi/(n-1)$, accept $H(i)$ (for $i \geq 2$).

(iv) If $p(2) \leq \pi/(n-1)$, reject $H(2)$ and move onto $H(3)$

(v) Proceed with the hypotheses until the first j such that $p(j) > \pi/(n-j+1)$

We can also apply Holm's method using Equation 4a ($\alpha = 1-(1-\pi)^{1/n}$, the Dunn-Ŝidák correction), in place of $\alpha = \pi/n$.

Simes-Hochberg Method

With Holm's method, we stop once we fail to reject a hypothesis. An improvement on this approach is the **Simes-Hochberg correction** (Simes 1986, Hochberg 1988), which effectively starts backwards, working with the largest p values first.

(i) If $p(n) \leq \pi$, then all hypothesis are rejected.

(ii) If not, $H(n)$ cannot be rejected, and we next examine $H(n-1)$.

(iii) If $p(n-1) \leq \pi/2$ then all $H(i)$ for $i \leq n-1$ are rejected.

(iv) If not, $H(n-1)$ cannot be rejected, and we compare $p(n-2)$ with $\pi/3$.

(v) In general, if $p(n-i) \leq \pi/(n-i+1)$ then all $H(i)$ for $i \leq n-i$ are rejected.

While the Simes-Hochberg approach is more powerful than Holm's, it is only strictly applicable when the tests within a family are independent. Holm's approach does not have this restriction. Hence, use Holm's if you are concerned about potential dependencies between tests, while if the tests are independent, use Simes-Hochberg or Hommel's method.

Hommel's Method

Hommel's (1988) method is slightly more complicated, but is more powerful than the Simes-Hochberg correction (Hommel 1989). Under Hommel's method, we reject all hypotheses whose p values are less than or equal to π/k, where

$$k = \max_i p(n-i+j) > \pi \, \frac{j}{i} \quad \text{for } j = 1, \cdots, i$$

Example 4 shows how all three of these methods are applied.

Example 4. Suppose for $n = 10$ tests, the (ordered) p values are as follows

i	1	2	3	4	5	6	7	8	9	10
$p(i)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{n-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

For an experiment-wide level of significance of $\pi = 0.05$, the Bonferroni correction is $\alpha = 0.05/10 = 0.005$. Hence, using a strict Bonferrioni for all, we reject hypotheses 1 and 2, and fail to reject (i.e., accept) 3-10. To applied the sequential methods, we use the associated $\alpha/(n-i+1)$ values under $\pi = 0.05$ which are also given in the table.

Under Holm's method, $p(i) \leq \pi/(n-i+1)$ for $i \leq 3$, and hence we reject $H(1)$ to $H(3)$ and accept the others.

Under Simes-Hochberg, we fail to reject $H(7)$ to $H(10)$ (as $p(i) > \pi/(n-i+1)$), but note that since $p(6) = 0.009 \leq \alpha/(n-i+1) = 0.010$, and hence we reject $H(6)$ to $H(1)$,

To apply Hommel's method, reject all hypotheses whose p values are less than or equal to π/k, where

$$k = \max_i p(n-i+j) > \pi \, \frac{j}{i}$$

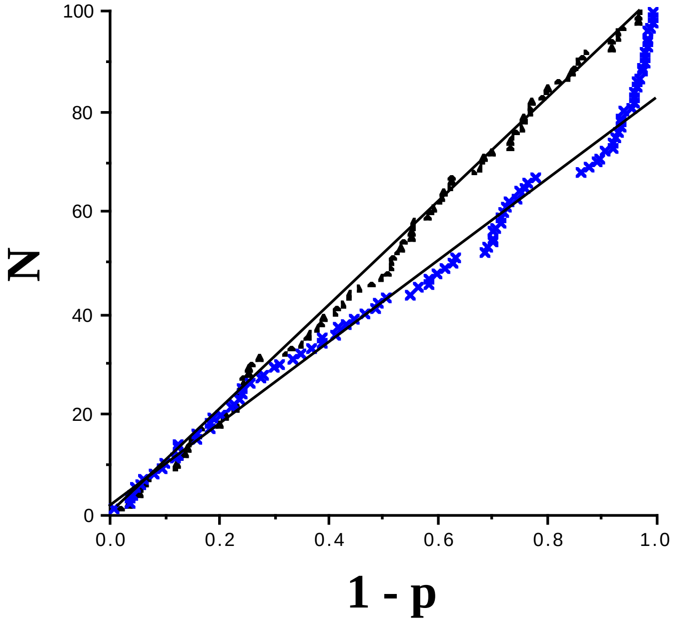
Let's start with $i = 1$. Here, (i=1, j=1), $p(10) = 0.5350 > \pi = 0.05$. Now lets try $i = 2$, giving (for $j = 1, 2$), $p(9) = 0.1055 > \pi(1/2) = 0.025$ and $p(10) > \pi$. For $i = 3$, $p(8) = 0.025 > \pi(1/3) = 0.0167$, $p(9) > \pi(2/3) = 0.033$, $p(10) > \pi$. For $i = 4$, $p(7) = 0.175 > \pi(1/4) = 0.0125$, but $(i = 4, j = 2)$, $p(8) = 0.025 = \pi(1/2)$. Hence, $k = 3$, and we reject all hypotheses whose p values are $\leq 0.05/3 = 0.0167$, which are $H(1)$ to $H(6)$. Note that a strict Bonferroni declared the fewest, and Simes-Hochberg and Hommel's the most, of the hypotheses to be significant.

Schweder-Spjøtvoll plots

A powerful result on the distribution of p values if we have a large set of truly null hypotheses is as follows:

The distribution of p values under the null hypothesis follows a Uniform(0,1) distribution.

This useful result has been used by a number of workers. One creative use are the plots of Schweder and Spjøtvoll (1982). Here, one orders the $1-p$ values from the smallest to the largest and plots the $1-p$ values on the horizontal axis, and N on the vertical axis. For example, the first point is $(1-p_n, 1)$, the second point $(1-p_{n-1}, 2)$, $\cdots$, the nth point point $(1-p_1, n)$. If all of the p values are indeed generated from null hypotheses, these are draw from a uniform and the resulting plot will be a straight line (the small triangles in the figure below).



Conversely, if some of the p values are draw from hypotheses where the null is false, we expect an excess of small p values, and hence an overabundance of $1-p$ values near 1. In the figure above, the x marks were generated from 80 true nulls and 20 significant hypotheses. Note the strong up-turn near one. Schweder and Spjøtvoll suggest can use these plots to estimate the actual number of true null hypothesis. One fits the best straight line until the upturn near one appears, extrapolating this line to obtain the N value when $1-p=1$ estimates the number of true null hypotheses, n_0. From the figure, their approach gives a value very close to 80, the correct number of true nulls used to generate this example.

FDR: The False Discovery Rate

Benjamini and Hochberg (1995) introduced an important concept for multiple comparison that they called the **false discovery rate**, or **FDR**. The FDR is the fraction of false positives among all tests declared significant. The motivation for using the FDR is that we may be running a very large number of tests, with those being declared significant being subjected to further studies. Examples might include looking for differential expression over a huge set of genes on a microarray or mapping a large number of genetic markers associated with a trait of interest. The initial analysis takes a large number of candidates and produces a reduced set for further analysis. In such cases, we are more concerned with making sure all possible true alternatives are included in this reduced set, and we are willing to put up with some false positives to accomplish this. However, we also don't want to be completely swamped with false positives. The idea is that the statistical procedure results in a significant *enrichment* of differentially-expressed genes, controlling the fraction of false positives within this enriched setting by specifying a value δ for the FDR. Choosing an FDR of 5% means that (on average) 5% of the genes we picked as being significant are actually false positives. The flip side is that 95% of those genes declared significant do indeed have differential expression. Hence, screening genes with an FDR of 5% results in a significant enrichment of genes that are truly differentially expressed.

To formally motivate the FDR, suppose a total of n hypotheses are tested, S of which are judged significant (by the criteria being used for each test). If we had complete knowledge, we would know that n_0 of the hypotheses have the null true and $n_1 = n - n_0$ have the alternative true, and we might find that F of the true nulls were called significant while T of the alternative true were called significant,

	Called significant	Called not significant	Total
Null true	F	$n_0 - F$	n_0
Alternative true	T	$n_1 - T$	n_1
Total	S	$n - S$	n

For this experiment, the false discovery rate is the fraction of tests called significant that are actually true nulls, $FDR = F/S$. (The term **discovery** follows in that a significant result can be considered as a discovery for future work.) As a point of contrast, the normal type 1 error (which we can also call the **false positive rate**, or **FPR**), is the fraction of true nulls called significant, is F/n_0. Note the critical distinction between these two in that while the numerator of each is F, the denominators are considerably different, the total number of tests called significant (for FDR) vs. the number of hypotheses that are truly null (FPR).

Another way to see the distinction between the false positive and false discovery rates is to consider them as probability statements for a single test involving hypothesis i. For the FDR we condition on the test as being significant,

$$\text{FDR} = \Pr(i \text{ is truly null} \,|\, i \text{ is significant}) = \delta \tag{5a}$$

where for the false positive rate, we condition on the hypothesis being null,

$$\text{FPR} = \Pr(i \text{ is significant} \,|\, i \text{ is truly null}) = \alpha \tag{5b}$$

Table 1 reminds the reader of the various test parameters that arise when multiple comparisons are considered. We now show how these various parameters are related.

Table 1: Multiple Comparisons Parameters

Parameter	Definition
α	Comparison-wise Type one error (false positive)
β	Type two error (false negative), $1-\beta$ = power
π	Family-wide Type one error, $\Pr(F>0)=\pi$
δ	False discovery rate
π_0	Fraction of all hypotheses that are null
p	Probability of the test statistic under the null
$p(k)$	k-th smallest p value of the n tests

First, the relationship between α, π and F is as follows. Suppose we have set the false positive rate (i.e., the Type one error rate) as α. Such a p value threshold (i.e., called significant if $p \leq \alpha$) only guarantees that the expected number of false positives is bound above by $E[F] \leq \alpha \cdot n$. For n tests, a π level experiment-wide false positive error (setting $\alpha = \pi/n$, the Bonferroni correction) implies $\Pr(F \geq 1) \leq \pi$, i.e., the probability of at least one false positive is π. To show how α, β, π_0, and δ are related, we first need to introduce the concept of the posterior error rate.

Morton's Posterior Error Rate (PER) and the FDR

Fernando et al. (2004) and Manly et al. (2004) have noted that FDR measures are closely related to Morton's (1955) **Posterior Error Rate** (**PER**), originally introduced in the context of linkage analysis in humans. Morton's PER is simply the probability that a single significant test is a false positive,

$$PER = \Pr(F = 1 \,|\, S = n = 1) \tag{6}$$

The connection between the FDR and PER is that if we set the FDR to δ then the PER for a randomly-drawn significant test is also δ.

Framing tests in terms of the PER highlights the **screening paradox** (Manly et al. 2004), "type I error control may not lead to a suitably low PER". For example, we might choose $\alpha = 0.05$, but the PER may be much, much higher, so that a test declared significant may have a much larger probability than 5% of being a false-positive. The key is that since we are *conditioning on the test being significant* (as opposed to conditioning on *the hypothesis being a null*, as occurs with α), this could include either false positives or true positives, and the relative fractions of each (and hence the probability of a false positive) is a function of the single test parameters α and β and fraction of null hypotheses, π_0. To see this, apply Bayes' theorem,

$$\Pr(F = 1 \,|\, S = n = 1) = \frac{\Pr(\text{false positive} \mid \text{null true}) \cdot \Pr(\text{null})}{\Pr(S = n = 1)} \tag{7}$$

Consider the numerator first. Let $\pi_0 = n_0/n$ be the fraction of all hypotheses that are truly null. The probability that a null is called significant is just the type I error α, giving

$$\Pr(\text{false positive} \mid \text{null true}) \cdot \Pr(\text{null}) = \alpha \cdot \pi_0 \tag{8a}$$

Now, what is the probability that a single (randomly-chosen) test is declared significant? This event can occur because we pick a null hypothesis and have a type I error or because we pick an alternative hypothesis and avoid a type II error. For the later, the power is just T/n_1, the fraction of all alternatives called significant. Writing the power as $1-\beta$ (β being the type II error, the failure to reject an alternative hypothesis), the resulting probability that a single (randomly-draw) test is significant is just

$$\Pr(S = n = 1) = \alpha\pi_0 + (1-\beta)(1-\pi_0) \tag{8b}$$

Thus

$$PER = \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + (1-\beta) \cdot (1-\pi_0)} \tag{9a}$$

$$= \frac{1}{1 + \frac{(1-\beta)\cdot(1-\pi_0)}{\alpha \cdot \pi_0}} \tag{9b}$$

In Morton's original application, since there are 23 pairs of human chromosomes, he argued that two randomly-chosen genes had a $1/23 \simeq 0.05$ *prior probability of linkage*, i.e., $1-\pi_0 = 0.05$ and $\pi_0 = 0.95$. Assuming a type I error of $\alpha = 0.05$ and 80% power to detect linkage ($\beta = 0.20$), this would give a PER of

$$\frac{0.05 \cdot 0.95}{0.05 \cdot 0.95 + 0.80 \cdot 0.05} = 0.54$$

Hence with a type-one error control of α = 0.05%, a random test showing a significant result ($p \leq 0.05$) has a 54% chance of being a false-positives. This is because most of the hypotheses are expected to null — if we draw 1000 random pairs of loci, 950 are expected to be unlinked, and we expect 950 · 0.05 = 47.5 of these to show a false-positive. Conversely, only 50 are expected to be linked, and we would declare 50 · 0.80 = 40 of these to be significant, so that 47.5/87.5 of the significant results are due to false-positives.

Note that the type I error rate of a test and the PER for a significant test, which are often assumed to be the same, we actually very, very different. The PER depends on power of a test and the fraction of tests that are truly null, in addition to depending on the type I error. Manly et al. (2004) note that the PER is acceptably low only if $1-\pi_0$ (the fraction of alternative hypotheses) is well above α.

Example 5. Suppose we set $\alpha = 0.005$ for each test, and suppose that the resulting power is essentially 1 (i.e. $\beta \simeq 0$). Consider 5,000 tests under two different settings. First, suppose that the alternative is very rare, with n_1 = 1 (π_0 = 0.9998). Under this setting, we expect 4,999 · 0.005 = 24.995 false positives and one true positive ($1 \cdot (1-\beta) = 1$), giving the expected PER as

$$\text{PER} = \frac{24.995}{24.995 + 1} = 0.961$$

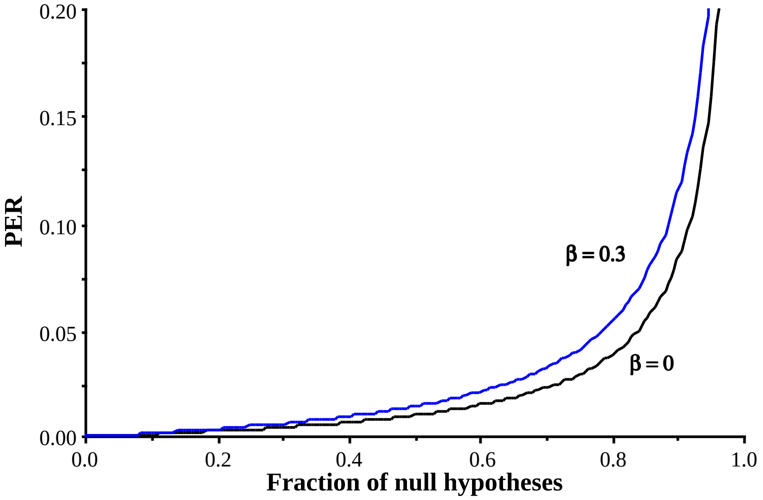
Thus a significant test has a 96.1% probability of being a false-positive.

Now suppose that the alternative is not especially rare, for example n_1 = 500 (π_0= 0.9). The expected number of false positives is 4500 · 0.005 = 22.5, while the expected number of true positives is 500, giving an PER of

$$\text{PER} = \frac{22.5}{522.5} = 0.043$$

The PER is thus rather sensitive to π_0, the fraction of all hypotheses which are null. If π_0 is essentially 1, an PER of δ is obtained using the Bonferroni correction, setting $\alpha = \delta/n$. However, if π_0 departs even slightly from one (i.e., more than a few of the hypotheses are correct), then the per-test level of α to achieve a desired PER rate is considerable larger than that given by the Bonferroni correction, i.e., $\alpha(\delta) > \delta/n$.

The figure below plots Equation (9b) assuming $\alpha = 0.0$ for various values of π (fraction of hypotheses that are null) and β (1-power).



Thinking in terms of the PER allows us to consider multiple comparisons in a continuum from Bonferroni-type corrections to using FDR to control the PER. If $\pi_1 = 1 - \pi_0$ is very small, most hypotheses tested are nulls and we wish to control the overall false positive rate with a Bonferroni-type correction. However, as some fraction of the hypotheses are expected to not be nulls ($1 - \pi_0$ is modest to large), then using FDR corrections makes more sense for controlling the PER.

A Technical Aside: Different Definitions of False Discovery Rate

While the false discovery rate for any experiment is just F/S, there are several subtly different ways to formally define the expectation of this ratio. The original notion of a false discovery rate is due to Benjamini and Hochberg (1995), with modifications suggested by a number of other workers, most notable Storey (2002) and Fernando et al (2004), see Table 2.

Table 2: Measures of False Discovery, (Manly et al. 2004)

	Name	Definition	Reference
FDR	False discovery rate	$E(\frac{F}{S} \mid S > 0)\,\Pr(S > 0)$	Benjamini and Hochberg (1995)
pFDR	Positive False discovery rate	$E(\frac{F}{S} \mid S > 0)$	Storey (2002)
PFP	Proportion of false positives	$E(F)/E(S)$	Fernando et al. (2004)
PER	Posterior error rate	$\Pr(F = 1 \mid S = n = 1)$	Morton (1955)
FPR	False Positive rate	$\Pr(F > 0)$	

While technically the distinction between these different false discovery rates is important, when actually estimating a false discovery rate from a collection of p values, one is usually left with an expression of the form $E(F)/E(S)$, the expected number of false positives to the expected number of significant tests. Strictly speaking, then, these are the proportion of false positives. This is a good thing, as Fernando et al. (2004) have shown that the PFP does not depend on either the number of tests or the correlation structure among tests (essentially this occurs because we are taking the ratio of two expectations, so the number of tests cancels in each and correlation structure among tests does not enter into the individual expectations).

The main operational differences between the different false discover rates are (i) the original method of Benjamini and Hochberg (1995), which assumes $n = n_0$ (all hypotheses are nulls), and (ii) all other estimators which assume n_0 is not necessarily one and thus also attempt to estimate either π_0 or n_0, and then uses these to estimate the false discovery rate.

The Original Benjamini-Hochberg FDR Estimator

The original estimate for the FDR was introduced by Benjamini and Hochberg (1995). Letting $p(k)$ denote the k-th smallest (out of n) of the p values, then the false-discovery rate δ_k for hypothesis k is bounded by

$$\frac{np(k)}{k} \leq \delta_k \tag{10a}$$

In particular, if we wish an FDR of δ for the entire experiment, then we reject (i.e., declare as significant) all hypotheses that satisfy

$$p(k) \leq \delta\,\frac{k}{n} \tag{10b}$$

Example 6. Consider again the 10 ordered p values from Example 4, and compute $n \cdot p(k)/k = 10p(k)/k$,

i	1	2	3	4	5	6	7	8	9	10
$p(i)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$10\frac{p(k)}{k}$	0.0200	0.0225	0.0200	0.0200	0.0170	0.0150	0.0250	0.0313	0.1172	0.5350

Thus, if we wish an overall FDR value of $\delta = 0.05$, we would reject hypotheses H(1) - H(8). Notice that this rejects more hypotheses than under any of the sequential Bonferonni methods (Example 4).

We will formally develop a more general estimate for the FDR below, but the basic idea leading to Equation 10a is as follows. Suppose we set a threshold value $p(k)$, declaring a test to be significant if its p value is at or below $p(k)$, in which case k of the hypotheses will be declared significant (as $p(k)$ is the k-th smallest p value), and $S = k$. Likewise, if all n of the hypotheses are null, then the expected value of F (the number of false positives) is just $n\,p(k)$. Thus the fraction of all rejected hypotheses that are false discoveries is just $F/S = n\,p(k)/k$, yielding Equation (10a).

This simple (heuristic) derivation shows while the original Benjamini-Hochberg estimate of the FDR is conservative, as in those settings in which one applies the FDR criteria, the expectation is that some fraction of the hypotheses are not null, and so $n_0 < n$. The correct estimate of the expected number of rejected null hypotheses is $n_0 p(k)$, leading to a more generalized estimate of the FDR as

$$\widehat{FDR} = \frac{\widehat{n_0}\,p(k)}{k} \tag{11}$$

where $\widehat{n_0}$ is an estimate of the number of truly null hypotheses out of the n being tested.

A (Slightly More) Formal Derviation of the Estimated FDR

Following Storey and Tibshirani (2003), consider the expected FDR for an experiment where we declare a hypothesis (or feature) to be significant if its p value is less than or equal to some threshold value, τ. Obviously, as τ becomes smaller, the FDR is smaller (as significant nulls become increasingly less likely). However, if τ is set too small, we lose power (e.g., suppose we set $\tau = \pi/n$, the Bonferonni correction). What we would like to do is to find the expected value of the FDR as a function of the chosen threshold τ to allow us to optimually tunde this parameter to control the desired FDR. With a large number of tested hypotheses,

$$E[FDR(\tau)] = E\left[\frac{F(\tau)}{S(\tau)}\right] \simeq \frac{E[F(\tau)]}{E[S(\tau)]} \tag{12}$$

A simple estimate of the expected number of significant tests when the threshold is set at τ is given by the observed number of significant tests when the threshold is τ.

To obtain an estimate for $E[F(\tau)]$, we call upon the property mentioned above that the distribution of p values under the null follows a uniform $(0,1)$ distribution. Hence,

$$\Pr(p \leq \tau \,|\, \text{null hypothesis}) = \int_0^\tau u(p)dp = \tau \tag{13a}$$

where $u(p)$ is the probability density function for p values under the null, which is just the uniform $(0,1)$ probability density function,

$$u(p) = \begin{cases} 1 & \text{for } 0 \leq p \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{13b}$$

Hence, if n_0 of the n tests are truly null, then

$$E[F(\tau)] = n_o \cdot \Pr(p \leq \tau \,|\, \text{null hypothesis}) \simeq n_0 \cdot \tau \tag{14}$$

Hence,

$$E[FDR(\tau)] = \frac{n_0 \cdot \tau}{S(\tau)} \tag{15}$$

Notice that we set $\tau = p(k)$, then $S(\tau) = k$, and Equation (14) become $n_0 p(k)/k$, recovering Equation (11).

Estimating the Number of Null Hypotheses, n_0

The problem remaining is to estimate n_0, the number of truly null hypotheses. Once again, we call upon the distribution of p values under the null being uniform. Recall that the histogram from a sufficiently large number of draws from this distribution is completely flat, as all values are equally likely. However, if some alternative hypotheses are mixed in with these nulls, then we expect the distribution to be a mixture, n_0/n being draws from a uniform and $1-n_0/n$ being draws from some other distribution in which the p values are skewed towards zero.

We have previously mentioned the regression estimator of Schweder-Spjøtvoll, but this tends to overestimate the number of nulls. Another approach was offered by Mosig et al. (2001), based on binning the p values. A third was suggested by Allison et al. (2002), who used ML to fit a mixture model to the p values, π_0 come from the null distribution (and hence a uniform), while $1 - \pi_0$ come from the alternative where p values are assumed to follow some (otherwise unspecified) beta distribution. The resulting likelihood functions

$$l(p) = (1-\pi_0)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1} + \pi_0 \tag{16}$$

The a and b parameters are fit by ML, as is the quantify of interest, π_0.

A very simple estimator was offered by Storey and Tibshirani (2003), using the key feature that draws from hypotheses which are not null are expected to have their p values skewed towards zero. Hence, if we look at the number of p values exceeding some tuning value λ (for example, $\lambda = 0.5$), then for large values of λ, most of these draws are from the uniform corresponding to draws from the null. We can use this fact to estimate $\pi_0 = n_0/n$ as follows. Let $\widehat{\pi_0}(\lambda)$ be the estimated based on using the tuning value λ, then

$$\widehat{\pi_0}(\lambda) = \frac{\text{Number of } p_i \text{ values } > \lambda}{n(1-\lambda)} \tag{17}$$

This follows from the uniform, as

$$\Pr(p > \lambda \,|\, \text{null hypothesis}) = \int_\lambda^1 u(p)dp = 1 - \lambda \tag{18}$$

where the density $u(p)$ for a uniform is given by Equation 13a. Hence, an estimate of n_0 is given by

$$\widehat{n}_0(\lambda) = n \cdot \widehat{\rho}(\lambda) = \frac{\text{Number of } p_i \text{ values } > \lambda}{1-\lambda} \tag{19}$$

Thus an estimated value for the FDR using threshold value τ (and based on tuning parameter λ) is just

$$\widehat{FDR}(\tau) = n_0 \cdot \frac{\tau}{S(\tau)} = \left(\frac{\text{Number of } p_i \text{ values } > \lambda}{1-\lambda}\right) \cdot \left(\frac{\tau}{\text{Number of } p_i \text{ values } \leq \tau}\right) \tag{20}$$

Ideally, over a reasonable range of λ values, we expect the estimate to ρ to be reasonably stable. If λ is set too large, the likelihood that almost all values correspond to draws from a null is countered by the much smaller sample size (and hence larger sampling error) from using such a small fraction of the total data.

Storey (*http://faculty.washington.edu/∼jstorey/qvalue*) has produced an **R** program, **Q-value** that (among other things) computes π_0 using this basic approach and then uses a smoother (a cubic spline) to average over different values of the tuning parameter. It also offers a bootstrap estimator for π_0.

Storey's q Value

While we can control the FDR for an entire set of experiments, we would also like to have an indication of the FDR for any particular experiment (or test) within this family. Intuitively, tests with smaller p values should also have smaller associated FDR values. Storey (2002, Storey and Tibshirani 2003) introduced the concept of a q value (as opposed to the p value) of any particular test, where q is the expected FDR rate for tests with p values at least as extreme as the test of interest. The estimated q value is a function of the p value for that test and the distribution of the entire set of p values from the family of tests being considered,

$$\widehat{q}\,[p(i)] = \min_{\tau \geq p(i)} \widehat{FDR}(\tau) \tag{21}$$

Example 7: As example of the interplay between the family-wide error rate π, and the individual p and q values for a particular test, consider Storey and Tibshirani 's (2003) analysis of a microarray data set from Hedenfalk et al. comparing BRCA1- and BRCA2 mutation positive tumors.

A total of 3,226 genes were examined. Setting a critical p value of $\alpha = 0.001$ detects 51 significant genes. (i.e., those with differential expression between the two types of tumors). Assuming the hypotheses being tested are independent (which is unlikely as expression is likely highly correlated across sets of genes), the probability of at least one false positive is $\pi = 1 - (1 - .0001)^{3226} = 0.96$, while the expected number of false-positives is 0.001· 3226 = 3.2, or 6% of the declared significant differences.

Setting a FDR rate of $\delta = 0.05$, Storey and Tibshirani detected 160 genes showing significant differences in expression. Of these 160, 8 (5%) are expected to be false-positives. Notice that, compared to the Bonferroni correction (51 genes, 6% false positives), over three times as many genes are detected, with a lower FDR rate. Further, Storey and Tibshirani estimate the fractionπ_0 of nulls (genes with no difference in expression) at 67%, so that 33% (or roughly 1000 of the 3226 genes) are likely differentially expressed.

To contrast the distinction between p and q values, consider the MSH2 gene, which has q value of 0.013 and p value of $5.50 \cdot 10^{-5}$. This p value implies that the probability of seeing at least this level of difference in expression given the null hypothesis (no difference in expression) is $5.50 \cdot 10^{-5}$. Conversely, $q = 0.013$ says that 1.3% of genes that show differences in expression that are as or more extreme (i.e., whose p values are at least as small) as that for MSH2 are false positives.

As a technical aside, why do we use $\min_{\tau \geq p(i)} \widehat{FDR}(\tau)$ instead of simply setting $q_i = \widehat{FDR}(p(i))$? Recall Example 6, where the original l Benjamini-Hochberg estimator for FDR value were used. This differs from other FDR estimators by a constant, n_0/n. Notice in particular that the smallest FDR occurs for hypothesis 6 (1.5%), and not for smaller p values. This reflects the tradeoff where increasing τ results in declaring more tests as significant, so that the ratio $\tau/S(\tau)$ need not monotonically increase as τ increases. As example 6 shows, setting the threshold τ *above* the $p(i)$ value may actually result in a smaller q value, and hence Storey's definition.

A final key point to stress above FDR methods is that all that is needed is the ordered listed of p values, with no other information about the testing really needed. Since the FDR rate is typically estimated using the PFP criteria (ratio of two expectations), the associated δ values are independent of the number of tests and their correlation structure. Stoery's aforementioned **`Q-value`** program takes an list of p values are returns the associated q values as well as estimates of π_0 and plots of q vs. p, the histogram of p values and other useful diagnostics. Note that the p value histogram should always be examined (akin to examining the plot of residuals in a fitted model). If the resulting p histogram is binomial, with modes near both zero and one, this indicates that at least some of the tests we likely one-sided, when two-sides tests are more appropriate.

References

Allison, D. B., G. L. Gadbury, M. Heo, J. R. Fernandez, C.-K. Lee, T. A. Prolla, and R. Weindruch. 2002. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* 39: 1-20.

Benjamini, Y., and Hochberg, T. 1995. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B* 85: 289–300.

Benjamini, Y., and Hochberg, T. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* 26: 60–83.

Fernando, R. L., D. Nettleton, B. R. Southey, J. C. M. Dekkers, M. F. Rothschild, and M. Soller. 2004. Controlling the proportion of false positives in multiple dependent tests. *Genetics* 166: 611-619.

Genovese C. and L. Wasserman. 2002. Operating Characteristics and Extensions of the False Discovery Rate Procedure. *Journal of the Royal Statistical Society Series B:* 64: 499–517.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800–802.

Holm, S. 1979. A simple sequential rejective multiple test procedure. *Scand. J. Statistics* 6: 65–70.

Hommel, G. 1988. A stagewise rejective multiple test procedure on a modified Bonferroni test. *Biometrika* 75: 383 – 386.

Hommel, G. 1989. A comparison of two modified Bonferonii procedures. *Biometrika* 76: 624-625.

Manly, K. F., D. Nettleton, and J. T. G. Hwang. 2004. Genomics, prior probability, and statistical tests of multiple hypotheses. (in press)

Morton, N. E. 1955. Sequential tests for the detection of linkage. *American Journal of Human Genetics* 7: 277–318.

Mosig, M. O., E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller, and A. Friedmann. 2001. A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Percentage in Israeli-Holstein Cattle, by Means of Selective Milk DNA Pooling in a Daughter Design, Using an Adjusted False Discovery Rate Criterion, *Genetics* 157: 1683-1698.

Schweder, T. and E. Spjøtvoll. 1982. Plots of p-values to evaluate many tests simultaneously. *Biometrika* 69: 493–502.

Simes, J. R. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 75–754.

Storey J.D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B:* 64: 479–498.

Storey J.D. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* 31: 2013-2035.

Storey J.D. 2004. QVALUE: The Manual (Version 1.0). On the web at *http://faculty.washington.edu/∼jstorey/qvalue/manual.pdf*

Storey J.D., J. E. Taylor, and D. Siegmund. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* 66: 187-205.

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100: 9440–9445.