Appendix 4

Multiple Comparisons: Bonferroni Corrections and False Discovery Rates

FDR methods commonly take a list of p-values and then determine for each data set how large the rejection threshold α can be made if we wish to keep the FDR below a bound
— Owen (2005)

Draft version 26 March 2014

Often one is faced with interpreting a list of p values from tests of hypotheses, either from a set of independent experiments all testing the *same* hypothesis or from a single experiment wherein a large number of *different* hypotheses are tested. Both of these are examples of **multiple comparisons**. In the former setting, the problem is to how best combine these independent p values into a single global statement of the evidence (or lack there of) in support of the common hypothesis. In the later setting, our concern is controlling error over the entire *collection* of tests from a single experiment, and this topic forms the bulk of this appendix.

Statistical analysis of a data set typically involves testing not just a single hypothesis, but rather many (often *very* many!). This is especially true in the genomics era wherein a single **high dimensional experiment** may test tens of thousands of hypotheses (such as treatment-dependent expression over all the genes in a genome). For any particular test, we may assign a preset probability α of a type I error (i.e., a **false positive**, declaring a test to be **significant**, namely $p \leq \alpha$, when in fact the null hypothesis is true). Under this broad setting, there are two different strategies for controlling error. First, if we expect that the vast majority of the tests likely follow the null, then we are interested in controlling the **experiment-wide error rate**, the probability of a false positive over *all* of the tests. The standard approach in this setting is the classic **Bonferroni correction** — obtaining an experiment-wide error rate of π over a set of n comparisons by declaring a test to be significant when $p \leq \alpha = \pi/n$. However, this is usually far too stringent and results in an enormous loss of power. We review sequential methods to improve this approach, but often the correction be best done by a shift in thinking. If we expect some reasonable number of the hypotheses to be false, then trying to avoid *any* false positives is not appropriate, but rather controlling the *fraction* of false positives in those tests we declare to be significant (**discoveries**) is a much better aim. This is especially true in large-scale exploratory experiments whose aim is to discover potential candidates for further studies. In this setting we attempt to control the **false discovery rate** (**FDR**), as opposed to the type I error (false positive) rate. Here the goal is to find a value τ such that the set of tests declared significant using $p \leq \tau$ has the desired false discovery rate.

Our treatment of these topics is as follows. First, we examine methods for combining p values over independent tests. We then turn to controlling the overall false-positive rate for a collection of tests from a single experiment through the use of Bonferroni corrections and their extensions. Given that the decision to control the false positives versus false discoveries hinges to a large extent on the fraction π_0 of the tests that are true nulls, we then examine how to estimate π_0 from the empirical distribution of the p values. We conclude by discussing approaches to control the false discover rate.

COMBINING p VALUES OVER INDEPENDENT TESTS

Hypotheses of interest are often testing in multiple studies, and an important issue (the statistical field of **meta-analysis** — the analysis of analyses) is how best to combine the results from these studies into a single global statement. The most obvious approach is simply to pool all the data and perform a single test, but for a variety of reasons this is often not feasible. For example, different tests of the same hypothesis may involve different methodologies and/or very different settings. Further, published papers may not report the full data set but rather just a few summary statistics. In such settings, one straightforward approach is to consider the list of p values for the collection of experiments that all purport to test the same hypothesis and try to obtain a single global p value for this entire set.

This simple question is potentially fraught with peril for several reasons. First, are the different tests all *really* testing the same hypothesis? The investigator must take care to assure this is correct before proceeding. Second, the so-called **file-draw effect**, wherein nonsignificant results remain in the file-draw (i.e., are not published), leading to published results being biased towards p small values. One general trend seems to be a publication bias for studies with small sample size but a reduction in this bias for larger samples (Easterbrook et al. 1991; Dickersin et al. 1992). The presumptive reason is that small studies often lack power, so that a nonsignificant result does not necessarily provide strong evidence that the null hypothesis is correct. Conversely, due to the higher power of larger studies, authors may feel more comfortable publishing negative results.

Fisher's χ^2

Fisher (1932) was among the first to offer a simple approach for combining p values (along with Tippett 1931), based on the important concept that *the distribution of p values under the null follows a uniform distribution over (0,1)*. Further, if $u \sim$ Uniform(0,1), then $-2\ln(u) \sim \chi^2_2$ (Pearson 1938). Hence, under the null, twice the negative natural log of a p value follows a chi-square distribution with two degrees of freedom. If we have k independent tests, then the sum of their log-transformed p values is the sum of k chi-square variables. Such a sum is itself chi-square with degrees of freedom given by the sum of the degrees of freedom for the individual chi-squares (LW Appendix 5). These observations lead to **Fisher's combined probability test**: for k independent tests with p_i denoting the p value for test i, the sum

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) \tag{A4.1}$$

approximately follows a χ^2_{2k} distribution.

Example A4.1. 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 ($\overline{p} = 0.09$). Fisher's statistic gives a value of

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

Hence, taken together these five tests show a highly significant p value. While the reader might find it surprising to obtain a significant value for the global p given that none of the individual

tests were, note that the distribution of p values is far from a uniform, but rather is skewed towards zero.

Rice (1990; also see Whitlock 2005) noted that a problem with Fisher's method is that small p values are differentially weighted compared to complementary large p values (e.g., p versus $1-p$). Equation A4.1 can be rearranged to yield

$$X^2 = -2k\ln(\overline{p}_G)$$

where $\overline{p}_G$ is the geometric mean of the individual p values, which differentially weights smaller values. Under Fisher's method an observed p value of (say) 0.001 receives more weight than a complementary value of 0.999, which is also as extreme. However, note for a unit normal U that $\Pr(U \leq -3.09) = 0.001$ while $\Pr(U \leq 3.09) = 0.999$, so that under a normal transformation the two complementary p values receive equal magnitude. This motivates the Z score method.

Stouffer's Z Score

A second approach for combining p values was offered by Stouffer et al. (1949), who transformed the individual p values into Z scores, obtained by solving $\Pr(U > Z) = p$. The sum of k independent unit normals is itself normal, with mean zero and variance k. These results lead to **Stouffer's Z score** method: for test i assign a score Z_i by solving $\Pr(U > Z_i) = p_i$, where U is a unit normal. Let Z_s denote the sum over the transformed p values of k tests, scaled by $k^{-1/2}$ to give the sum a variance of one,

$$Z_s = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{A4.2a}$$

Since $Z_s \sim N(0,1)$, the overall p value is obtained by

$$p = \Pr(U > Z_s). \tag{A4.2b}$$

As noted by Whitlock, this test was first proposed in a footnote in the authors' sociological study of Army life, making it one of the more obscure origins of a statistical method!

Example A4.2. Reconsider the data from Example A4.1. The Z_i value are easily obtained using **R**, as the command **`qnorm(1-p)`** returns Z satisfy $\Pr(U \leq Z) = 1-p$, or (equivalently) that $\Pr(U > Z) = p$. For example, Z_1 is given by **`qnorm(1-0.1)`**, or 1.281. Similarly computing the other Z_i values gives

$$\sum_{i=1}^{5} Z_i = 6.754, \quad \text{hence} \quad Z_s = \frac{6.754}{\sqrt{5}} = 3.020$$

Since $\Pr(U > 3.020) = 0.00126$, as in Example A4.1, the combined p value is highly significant.

Besides providing symmetric values for large and small p values (i.e., p and $1-p$), a second major advantage of the Z score approach is that one can individually *weight* p values

from different tests (Mosterller and Bush 1954; Liptak 1958), as the weighted sum of normals is itself a normal (while the weighted sum of χ^2 variables — the analog for Fisher's test — is considerably more complex). The resulting weighted version becomes

$$Z_w = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \tag{A4.2b}$$

where $Z_w \sim N(0,1)$. As expected, Z_w reduces to Z_s when all the weights are equal. One can either weight by the degrees of freedom or by the reciprocal of the standard error of the estimate. Whitlock (2005) shows that the weighted Z score method is superior to either X^2 or Z_s when sample size various over data. Z_w has higher power and also a higher correlation between its predicted p value and the actual p value obtained if one was able to merge all the samples. As noted by Whitlock (2005), many studies in evolutionary biology are interested in whether a hypothesis consistently holds over a collection of species. In such cases, the number of species is number of replicates, and weighting p values for individual species is inappropriate.

BONFERRONI CORRECTIONS AND THEIR EXTENSIONS

We now turn to the complementary problem of determining the significance level α for individual tests required to control the overall false positive rate over a collection of n tests. The typical setting is that a single study or experiment has gathered data and a number of different tests, usually on *different* hypotheses, are performed using this data. Let π denote our desired experiment-wide false positive rate — the probability of one (or more) false positives *over the entire collection* of n tests is no greater than π. The standard approach for determining the appropriate α given n and π is to use **Bonferroni corrections**.

Standard Bonferroni Corrections

The probability of not making any type I errors (false positives) over n independent tests, each at level α, is $(1-\alpha)^n$. Hence, the probability of at least one false positive over the entire collection is just one minus this,

$$\pi = 1 - (1-\alpha)^n \tag{A4.3a}$$

Solving for the α value required for each test gives

$$\alpha = 1 - (1-\pi)^{1/n} \tag{A4.3b}$$

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

$$\alpha = \pi/n \tag{A4.4}$$

Both Equations A4.3b and A4.4 are referred to as Bonferroni corrections. In the literature, π is the **family-wide error rate** (**FWER**), while α is the **comparison-wise error rate**, or **CWER**.

Example A4.3. Suppose we have $n = 100$ independent tests and wish an overall π value of 0.05. What α should be used for each individual test to achieve an experimental-wide false positive rate of 0.05? 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 gives $\overline{x} \pm 3.48\sqrt{\text{Var}}$.

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 **sequential Bonferroni corrections**. Sequential approaches have increased power over standard Bonferroni corrections, as is illustrated (below) in Example A4.4. Shaffer (1995) reviews these, and other, approaches.

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 declared 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 rejecting hypotheses until the first i such that $p(i) > \pi/(n-i+1)$

We can also apply Holm's method using Equation A4.3a ($\alpha = 1-(1-\pi)^{1/n}$, 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 A4.4 shows how all three of these methods are applied.

Example A4.4. 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 apply these sequential methods, we use the associated $\pi/(n-i+1)$ values for $\pi = 0.05$ which are 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 since $p(6) = 0.009 \leq \pi/(n-i+1) = 0.010$, 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), $p(10) = 0.5350 > \pi \cdot (1/1) = 0.05$. Now lets try $i = 2$, giving (for $j = 1, 2$), $p(9) = 0.1055 > \pi(1/2) = 0.025$ and (as above) $p(10) > \pi$. For $i = 3$, $p(8) = 0.025 > \pi \cdot (1/3) = 0.0167$, $p(9) > \pi \cdot (2/3) = 0.033$, $p(10) > \pi$. For $i = 4$, $p(7) = 0.175 > \pi \cdot (1/4) = 0.0125$, but ($i = 4, j = 2$), $p(8) = 0.025 = \pi \cdot (1/2)$. 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 while controlling the experiment-wide false positive rate at 0.05.

Dealing with Dependence: The Leek-Storey Surrogate Variable Approach

What happens when p values are *not* independent? Indeed, how does one even know when a lack of independence exists among the hypotheses tested? In a typical high-dimensional experiment, a series of n variables are measured in m separate settings (typically individuals) to give m data vectors, where $n \gg m$, often by orders of magnitude. For example, a standard gene expression study may follow the mRNA levels of tens of thousands of genes in just a few dozen individuals. In such settings where the number of hypotheses to be tested is orders of magnitude greater than the number of individual data vectors, dependency among tests is guaranteed. The critical observation, as noted by Leek and Storey (2008), is

that this dependency can be removed *before* individual tests are performed, and that removal of this dependency generates independent p values among the resulting tests. While one can (potentially) control for lack of independence from the covariance structure among the tests, this is of order n^2 (one must use roughly n^2 variance and covariance terms to account for this, Owen 2005), while the covariance structure among the data is (at most) of order m^2. Leek and Storey (2007, 2008) discuss how to estimate this common latent structure in the data and remove it before individual testing, an approach they call **surrogate variable analysis**. We discuss a similar problem in Volume 3, the decomposition of a G x E matrix into a lower-dimensional approximation using the singular value decomposition.

DETECTING AN EXCESS OF SIGNIFICANT TESTS

While Bonferroni corrections (and their sequential counterparts) are widely used, when the number of tests is modest to large their application significantly erodes power, leading to very high Type II errors (failing to declare a test significant when the null is false). This Type I versus Type II tradeoff applies to most statistical tests (LW Appendix 5) and which error is more of a concern to the investigator determines how to proceed. In many cases, our initial experiment is simply an enrichment method: we wish to take a large number of possible hypotheses and extract a subset showing the most support for their alternative hypotheses for further consideration. In such cases, we are often more concerned with Type-II errors, as the penalty of including a test that is from the null may be less than the penalty of excluding a test that is not from the null. In such settings, we would like to calculate the number n_0 of tests (or equivalently the fraction $\pi_0 = n_0/n$) of true nulls among the n tests. The first step towards doing so is to ask if the observed number of significant tests is excessive under the **global null hypotheses** (all tests are from true nulls).

How Many False Positives?

Suppose we perform n independent tests, each with a type I error rate α. If all tests are from their respective null hypotheses, the number j of false positives is Binomial with α the probability of a "success" (a false positive) and n the number of trials,

$$\Pr(j \text{ false positives}) = \frac{n!}{(n-j)!\,j!}(1-\alpha)^{n-j}\,\alpha^j \tag{A4.5}$$

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

$$\Pr(j \text{ false positives}) \simeq \frac{(n\alpha)^j e^{-n\alpha}}{j!} \tag{A4.6}$$

Example A4.5. Suppose 250 independent tests are performed, each with $\alpha = 0.025$ (a 2.5% chance of declaring a result from the null hypothesis to be significant), and 15 tests are declared significant by this criteria. Is this number greater than expected by chance? The expected number of significant tests under the global null hypothesis is $n\alpha = 250 \cdot 0.025 = 6.25$. From Equation A4.5 the probability of observing 15 (or more) significant tests is

$$\sum_{j=15}^{250} \Pr(j \text{ false positives}) = \sum_{j=15}^{250} \frac{250!}{(250-j)!\,j!}(1-0.025)^{250-j}\,0.025^j$$

We could either sum this series directly or use the cumulative distribution function for a binomial, which is many statistical packages. In **R**, the probability that a binomial with parameters n and p has a value of i or less is obtained by **`pbinom(i,n,p)`**. The probability of 15 or greater is one minus the probability of 14 or less, or **`1- pbinom(14,250,0.025)`**. **R** returns 0.00177. Given that there is only a 0.2% of seeing this many significant tests under the global null, we expect some of these significant tests to be true **discoveries** (the null hypothesis is incorrect), not false positives. The critical question, of course, is which ones?

Testing for an excessive number of significant tests is a rather crude indicator of the actual number n_0 of the n tests that are true nulls. It is very possible that $n_0 < n$ and yet we would not detect an excess of significant tests by the above method. Likewise, if an excessive number is detected, what really can we say about n_0 other than $n_0 < n$? For instance, Example A4.5 shows an excess of 9 significant tests (observed 15, expected 6), but clearly assuming $n_0 = n - 9$ is a bit naive. Finally, the outcome varies with our choice of α. One could easily imagine an excess of significant tests using $\alpha = 0.05$, but not when using $\alpha = 0.01$. Ideally, we would like to have an estimate for n_0 that is independent of the choice of α.

Such estimators readily follow from the key idea in this Appendix, namely that if the null is correct, draws of p values follow a uniform over (0,1). A more careful examination of the empirical distribution of p values over our sample of tests, rather than simply how many we declare significant, is the key to obtaining estimates of n_0.

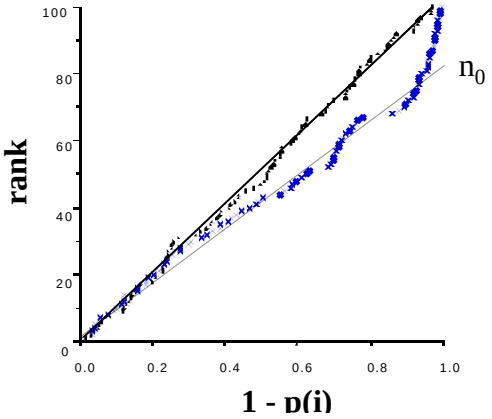

Figure A4.1: A Schweder-Spjøtvoll plot is one approach for detecting departures from a uniform distribution of p values. The p values are ordered from smallest $p(1)$ to largest $p(n)$, and one plots their rank as a function of $1 - p$. Under a uniform, the result is a straight line passing through the origin and the point $(1, n)$. The upper curve, generated by randomly sampling all $n = 100$ values from a uniform (0,1), fits the pattern. The lower curve, generated by simulating p values for 80 true nulls and 20 tests where the alternative was correct, shows an inflation of p values near zero ($1 - p$ values near one). This results in a strong departure from linearity near one. Ignoring this upturn and extrapolating the linear fit for the values below this inflection point gives an approximate value of 80 for $1 - p = 1$, which is our estimate of n_0.

Schweder-Spjøtvoll plots

A simple graphical approach using the empirical distribution of p values was suggested by

Schweder and Spjøtvoll (1982). If one rank-orders the p values from the smallest $p(1)$ to the largest $p(n)$, a plot of $p(i)$ versus i is a straight line under a uniform. Since our interest is usually in detecting an excessive number of small p values (as would be expected if $n_0 < n$), Schweder and Spjøtvoll suggest plotting $1 - p(i)$ values on the horizontal axis, and ranks of these [which are the reverse of the ranks of the $p(i)$] on the vertical axis. For example, the first point is $(1 - p(n), 1)$, the second $(1 - p(n-1), 2)$, $\cdots$, and the nth $(1 - p(1), n)$. 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 (upper curve in Figure A4.1). 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 one (lower curve in Figure A4.1).

In addition to providing a quick visual check as to whether the p values follow a uniform, Schweder and Spjøtvoll suggest that these plots can also estimate n_0. One fits the best straight line until the upturn near one appears, extrapolating this line to obtain the n value for $1-p = 1$ estimates the number of true null hypotheses, n_0. As shown in Figure A4.1, this gives a value very close to 80, the correct number of true nulls used to generate this example.

Estimating n_o: Subsampling from a Uniform Distribution

As suggested by the Schweder-Spjøtvoll plot, the distribution of p values offers insight into the number of truly null hypotheses n_0. While this plot offers either a simple visual, or a more formal regression-based, estimator of n_0, it tends to overestimate the number of nulls. A number of other estimators have been suggested, again based on the distribution of p values for those tests under the null being uniform. Recall that the histogram from a sufficiently large number of draws from this distribution is flat, as all values are equally likely (Figure A4.2A). However, if the null is false for all tests, then the distribution of p values is shifted away from uniform, usually with a skew towards smaller values (Figure A4.2B), but potentially also skewed towards one (for example, if one-tailed tests are used when a two-tailed test is appropriate, Figure A4.2C).

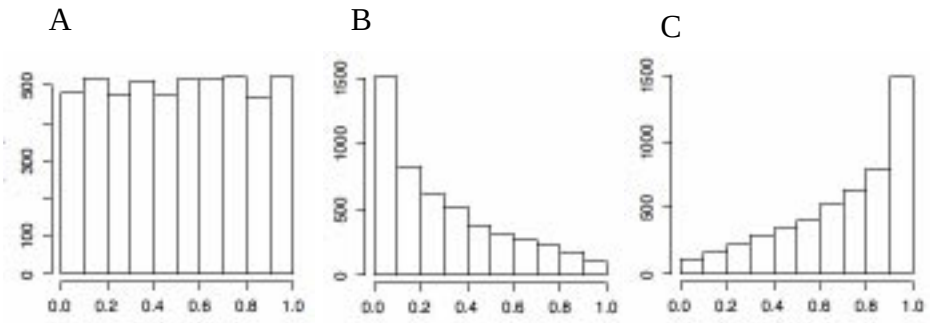

Figure A4.2: Simulated distribution of p values based on 5000 tests for samples of 25 draws from a normal distribution with mean μ and variance one. The null hypothesis is $H_0 : \mu \leq 0$. **A:** The distribution of p values when $\mu = 0$ (the null is correct) is uniform. **B:** The distribution when $\mu = 0.2$ is skewed towards an excess of values near zero. **C:** The distribution when $\mu = -0.2$ is skewed towards an excess of values near one.

If the collection of tests contains some alternative hypotheses mixed in with true nulls, we expect the distribution to be a mixture, with fraction $\pi_0 = n_0/n$ being draws from a uniform and $(1 - n_0/n)$ from some other distribution. Figure A4.3 plots the empirical distribution of p values from a study by Mosig et al. (2001) on marker-trait associations. While the middle of the distribution appears to be consistent with random sampling around

a flat average, there is a large excessive of values near zero.

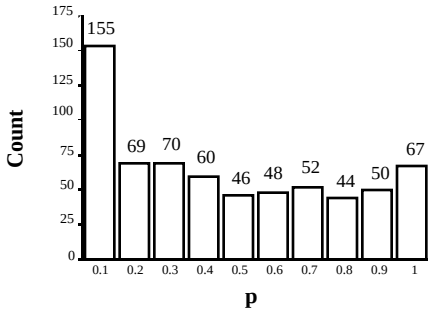

Figure A4.3: An empirical distribution of p values, from a study by Mosig et al. (2001). The number of p values in each of ten bins (of length 0.1) are given above the bars. Note the large excess of values near zero.

One simple approach for estimating n_0 is to use the average height for middle values in the p-value histogram. Presumably, these are almost entirely drawn from null hypotheses, while this may not be the case for values near zero (and potentially one). Recall that the probability density function ϕ_u for a uniform over (0,1) has a very simple form,

$$\phi_u(p) = \begin{cases} 1 & \text{for } 0 \le p \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{A4.7a}$$

If there are n_0 truly null tests, then the expected number of p values from these tests falling within an interval $0 \le a < b \le 1$ is just

$$n_0 \int_a^b \phi_u(p)\, dp = n_0 \int_a^b 1 \cdot dp = n_0(b-a) \tag{A4.7b}$$

Hence

$$\widehat{n}_0(a,b) = \frac{\text{Number of } p(i) \text{ values in } (a,b)}{b-a} \tag{A4.7c}$$

Likewise, an estimate for the fraction $\pi_0 = n_0/n$ of true nulls is

$$\widehat{\pi}_0(a,b) = \frac{\text{Number of } p(i) \text{ values in } (a,b)}{n(b-a)} \tag{A6.7c}$$

$$= \frac{\text{fraction of } p(i) \text{ values in } (a,b)}{b-a} \tag{A6.7d}$$

Example A4.5. Using the data in Figure A4.3, what is n_0? Consider bins centered around $p = 0.05$. Using the three binds 0.4, 0.5, and 0.6, a total of 60 + 46 + 48 = 155 tests have p values in this interval. From Equation A4.7b, $155 = n_0 \cdot 0.3$ or $n_0 = 155/0.3 = 516$, and hence a fraction $\pi_0 = n_0/n = 516/644 = 0.80$ of the tests are true nulls. Using the bins 0.3 to 0.8 gives $n_0 = 322/0.6 = 537$, giving $\pi_0 = 0.83$. Hence, it appears that around 80% of the tests are consistent with true nulls. Mosig et al. (2001; also see Nettleton et al. 2006) using an

iterative approach (also based on bin counts in the p-value histogram) arrived at an estimate of $n_0 = 500$ (78%).

Storey and Tibshirani (2003) considered the number of p values exceeding some tuning value λ (so that $a = \lambda$ and $b = 1$ in Equation A4.7b). Their logic being that for large values of λ, most of these draws are from the uniform corresponding to draws from the null. Let $\widehat{\pi}_0(\lambda)$ denote the estimated based on using the tuning value λ, then

$$\widehat{\pi}_0(\lambda) = \frac{\text{Number of } p(i) \text{ values } > \lambda}{n(1-\lambda)} \tag{A4.8a}$$

and

$$\widehat{n}_0(\lambda) = n \cdot \widehat{\pi}_0(\lambda) = \frac{\text{Number of } p(i) \text{ values } > \lambda}{1-\lambda} \tag{A4.8b}$$

By focusing on the interval $(\lambda, 1)$, the **Storey-Tibshirani estimator** is potentially biased when there are an excess of p values near one. This can happen for a variety of reasons, such as inappropriate assumptions for the test statistic (e.g., the use of one-sided tests when two-sided tests are more appropriate). Both Equation A4.7c and the Storey-Tibshirani estimator (Equation A4.8b) rely on tuning parameters (a, b and λ) which define the region of the distribution of p values assumed to be drawn from a uniform (i.e., almost all p values in this interval are assumed to be generated under the null). Nettleton et al. (2006) reviews these and other approaches for estimating n_0 from sampling parts of a presumed uniform and compares their strengths and weaknesses.

One significant concern is that when tests are correlated, this can result in either an under- or over-dispersion of p values under the global null hypothesis, resulting in significant departure from a uniform distribution (Efron 2007, Hu et al. 2011, Leek and Storey 2011). This in turn compromise estimates of n_0. Corrections for dealing with correlated tests have been proposed (e.g., Owen 2005; Efron 2007; Leek and Storey 2007, 2008). As discussed earlier, the approach of Leek and Storey for first accounting for dependence in the data before computing individual p values seems the most promising and results in a uniform distribution of p values under the global null.

Estimating n_o: Mixture Models

Allison et al. (2002) suggested that π_0 can be estimated by treating the distribution of p values as a mixture, π_0 of which come from a uniform (and hence distribution function ϕ_u) while the remainder $(1-\pi_0)$ are from the distribution $\phi_A(p)$ of p values when the alternative hypothesis is true. While the general form of $\phi_A(p)$ is unknown, a very flexible distribution to model it is the beta (Appendix 2),

$$\phi_A(p) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, p^{a-1}(1-p)^{b-1} \tag{A4.9a}$$

Under the alternative, we expect an increase in p values near zero, which occurs when $a < 1$. Likewise, the beta can easily accommodate an increase in p values near one as well ($b < 1$). When $a = b = 1$, this simply reduces to a uniform. Allison et al. suggested to fit the actual shape by using the data to obtain ML estimates of a and b, as well as our desired parameter π_0. The resulting likelihood function for a single p value becomes

$$\ell(p) = (1-\pi_0)\,\phi_A(p) + \pi_0\,\phi_u(p) = (1-\pi_0)\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, p^{a-1}(1-p)^{b-1} + \pi_0, \tag{A4.9b}$$

with the resulting total likelihood over n p values (from independent tests) becoming

$$\ell(\mathbf{p}) = \prod_{i=1}^{n} \ell(p_i) \qquad \text{(A4.9c)}$$

Standard ML methods (LW Appendix 2) are used to solve for a, b, π_0. More generally, one can fit k separate beta distributions for $\phi_A(p)$, using ML to estimate the mixture proportions and parameter values, see Allison et al. (2002) for details.

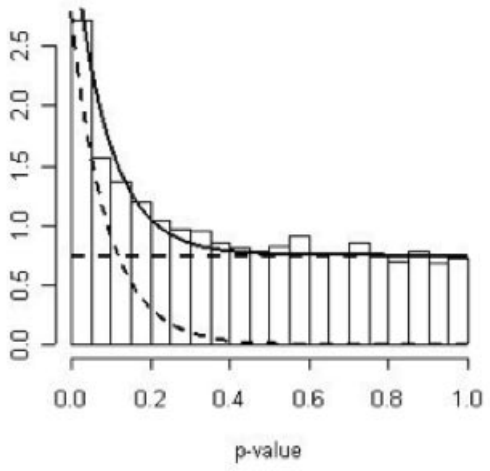

Figure A4.4: The empirical distribution of p values can be treated as a mixture model of a uniform plus a beta (whose shape parameters a and b can be estimated via ML), see Equation A4.9b. In this hypothetical example, a uniform (horizontal dashed line) and a beta with ($a < 1, b = 1$), dashed curve, when weighted yield the mixture distribution (solid curve) that fits the empirical distribution of the p values.

While normally hypotheses testing under a maximum likelihood framework is done using the likelihood ratio (LR) test (LW Appendix 2), this is not appropriate for tests of the number of components in a mixture, as LR does not approach a limiting χ^2 distribution (McLachlan 1987). While a modified LR tests for mixtures can be constructed that is better behaved (Chen et al. 2001), Allison et al. used a bootstrap approach (McLachlan 1987; Schork 1992). Here, one first uses the original distribution of p values to compute a LR test statistic for the null of a uniform versus the alternative of a mixture. One then generates **parametric bootstrap samples** by drawing n p values from the null distribution (here a uniform) and then using this simulated dataset to compute a LR test statistic for a mixture. This is done several thousand times to generate an approximate distribution of the LR statistic under the null, which is used to assess significance. For example, if only 0.25% of the bootstrap LR values are equal to (or exceed) the LR value for the original data, the significance is approximately 0.25%. Likewise, approximate standard errors for π_0 can be generated using a **conventional bootstrap** approach. One samples the *original* p values with replacement to generate a **bootstrap sample** of size n. This is used to estimate π_0 (and the other parameters) under a standard ML framework. A thousand or more bootstrap samples are generated, and the variation across estimates of π_0 (or any other parameter) over these samples provides an approximate estimate of the sampling variance.

Finally, while a beta (or weighted sum of betas) can be used as the functional form for ϕ_A, another approach is to use a non-parametric estimator for this unknown density function.

This can be done using a **kernel density estimator**, where the form of an unknown density is estimated by using the observed number of counts within a series of bins spanning the distribution in conjunction with an appropriate smoothing function. This approach has been used by Robin et al. (2007) and Guedj et al. (2009).

FDR: THE FALSE DISCOVERY RATE

We can loosely group issues of multiple comparisons into three problems, two of which have been discussed: combining p values from independent tests of the same hypothesis and control of the overall false positive rate for a collection of different tests from the same experiment (Bonferroni corrections and their extensions). Such corrections are appropriate when we expect only a few of the many hypotheses being false. An alternate setting is that some substantial fraction of the tests are indeed expected to be false. In such cases, even sequential Bonferroni correction are likely too stringent, resulting in too many false negatives (Type II errors, failure to reject a false hypothesis). A different approach is required in these settings, and this is the **false discovery rate**, or **FDR**, introduced by Benjamini and Hochberg (1995).

The FDR is the fraction of false positives among all tests declared significant. The motivation for using the FDR is that we may be conducting a very large number of tests, with those being declared significant being subjected to further studies. An example would be searching for differential expression over a huge set of genes on a microarray. The goal of the initial analysis is to take a large number of candidates and distill a reduced set for further analysis that is highly enriched for true positives. 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 accept some false positives to accomplish this goal. 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 true positives (differentially-expressed genes in our example), while controlling the fraction of false positives within this enriched set by specifying a value δ for the FDR. Choosing an FDR of 5% means that (on average) 5% of the genes we declare as being significant are actually false positives. The flip side is that 95% of those genes (tests) 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 (the p value for that test is less than or equal to some threshold value τ). 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 I 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 S of tests called significant (for FDR) versus the number n_0 of hypotheses that are truly null (FPR). As the threshold value τ is changed, so is F/S. To obtain an FDR of δ over our experiment,

τ is adjusted to find its largest value such that some expectation of F/S is bounded above by δ. Finally, Gadbury et al. (2004) define the **expected discovery rate** (**EDR**) as T/n_1 (the fraction of all true discoveries declared as significant), which is the analogue of statistical power in this setting.

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,

$$\text{FDR} = \Pr(i \text{ is truly null} \,|\, i \text{ is significant}) = \delta \tag{A4.10a}$$

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

$$\text{FPR} = \Pr(i \text{ is significant} \,|\, i \text{ is truly null}) = \alpha \tag{A4.10b}$$

Table A4.1. Summary of the multiple comparisons parameters used in this Appendix. F denotes the number of false positives — tests under the null that are declared significant.

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

Table A4.1 reminds the reader of the various test parameters that arise when multiple comparisons are considered. We now show how these various parameters are related. The relationship between α, π, and F is as follows. Suppose we have set the false positive rate (i.e., the Type I error rate) for an individual test at α. Such a p value threshold only guarantees that the expected number of false positives is bounded above by $E[F] \leq \alpha \cdot n$. For n independent 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,

$$PER = \Pr(F = 1 \,|\, S = n = 1) \tag{A4.11}$$

The connection between 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), S 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 π_0 of null hypotheses. To see this, apply Bayes' theorem (Equation A2.1),

$$\Pr(F = 1 \,|\, S = n = 1) = \frac{\Pr(\text{false positive} \,|\, \text{null true}) \cdot \Pr(\text{null})}{\Pr(S = n = 1)} \tag{A4.12}$$

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 declared significant is just the type I error α, giving

$$\Pr(\text{false positive} \,|\, \text{null true}) \cdot \Pr(\text{null}) = \alpha \cdot \pi_0 \tag{A4.13a}$$

Now, what is the probability that a single (randomly-chosen) test is declared significant? This event can occur because we pick a null hypothesis (π_0) and have a type I error (α) or because we pick an alternative hypothesis ($1 - \pi_0$) and avoid a type II error. For the later, the power is just T/n_1, the fraction of all alternatives called significant. Writing power as $1 - \beta$ (β being the type II error), the resulting probability that a single (randomly-draw) test is significant is just

$$\Pr(S = n = 1) = \alpha\pi_0 + (1 - \beta)(1 - \pi_0) \tag{A4.13b}$$

Thus,

$$PER = \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + (1 - \beta) \cdot (1 - \pi_0)} = \left(1 + \frac{(1 - \beta) \cdot (1 - \pi_0)}{\alpha \cdot \pi_0}\right)^{-1}. \tag{A4.14b}$$

Figure A4.5 plots for for various values of π_0 and β.

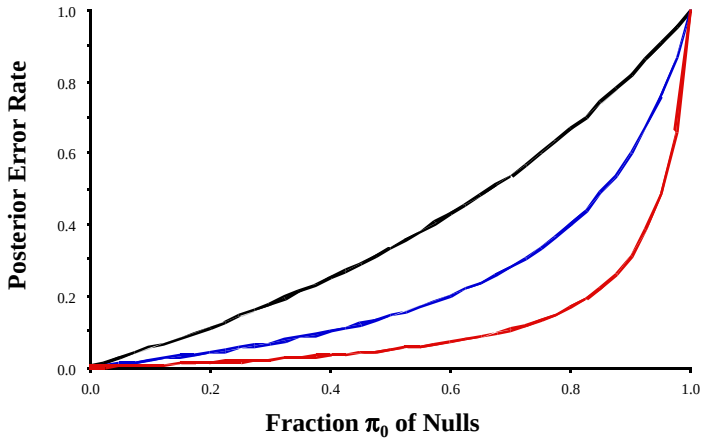

Figure A4.5. Plot of the posterior error rate (Equation A4.14) for $\alpha = 0.05$ as a function of the fraction π_0 of null hypotheses and the type II error β (one minus the power). Upper curve corresponds to $\beta = 0.9$ (10% power), middle to $\beta = 0.7$ (30% power), and lower curve to $\beta = 0$ (100% power).

Example A4.6. 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 I 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 occurs because most of the hypotheses are expected to be 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 = 0.54 of the significant results are due to false-positives.

What value for α is needed under the above parameters to given a PER of 0.05? Solving for α in

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

gives $\alpha = 0.0022$. Hence, setting this as the type I error gives a PER of five percent.

The type I error rate of a test and the PER for a significant test, which are often assumed to be the same, are actually very different. The PER is a function of the power of a test and the fraction of tests that are truly null, as well as 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 A4.7. 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, $\alpha = \delta/n$. 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 (i.e., less stringent) than that given by the Bonferroni correction, i.e., $\alpha(\delta) > \delta/n$. For example for a 0.04 experiment-wide error rate, $p = 0.04/5000 = 8 \times 10^{-6}$, roughly 625 times the value of $p = 0.005$ required for a 4% FDR, greatly increasing power under the FDR framework.

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 be false ($1-\pi_0$ 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 A4.2.

Table A4.2 Measures of false discovery, after 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, this is the **proportion of false positives** (**PFP**).

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 use these to estimate the false discovery rate.

The 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 of the p values, then the false-discovery rate δ_k for hypothesis k is bounded by

$$\frac{np(k)}{k} \leq \delta_k \tag{A4.15a}$$

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

$$p(k) \leq \delta \frac{k}{n} \tag{A4.15b}$$

Example A4.8. Consider again the 10 ordered p values from Example A4.4, and compute $n\,p(k)/k = 10\,p(k)/k$,

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 when $n\,p(k)/k \leq \delta = 0.05$, which are H(1) - H(8). Notice that this rejects more hypotheses than under any of the sequential Bonferonni methods (Example A4.4).

We formally develop a more general estimate for the FDR below, but the basic idea leading to Equation A4.15a 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 $\tau = p(k)$, 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 $n\,p(k)$. The resulting fraction of all rejected hypotheses that are false discoveries is just $F/S = n\,p(k)/k$, yielding Equation A4.15a.

This simple (heuristic) derivation shows why 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 estimator of the expected number of rejected null hypotheses is $n_0 p(k)$, leading to a more generalized estimate of the FDR where $\widehat{n}_0$ replaces n (e.g., Equations A4.7-A4.9). For example, Equation A4.15a becomes

$$\widehat{\delta}_k = \frac{\widehat{n}_0\,p(k)}{k} \tag{A4.16}$$

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 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 optimally tune this parameter to control the desired FDR. With a large number of tested hypotheses,

$$E[FDR(\tau)] = E\left[\frac{F(\tau)}{S(\tau)}\right] \simeq \frac{E[F(\tau)]}{E[S(\tau)]} \tag{A4.17}$$

A simple estimate of $E[S(\tau)]$ is given by the observed number of significant tests when the threshold is τ.

To obtain an estimate for $E[F(\tau)]$ we again use the distribution of p values under the null following a uniform $(0, 1)$ distribution. Hence,

$$\Pr(p \leq \tau \,|\, \text{null hypothesis}) = \int_0^\tau \phi_u(p)dp = \tau \tag{A4.18}$$

where $\phi_u(p)$ is the uniform probability density function for p values under the null (Equation A4.7a). Hence, if n_0 of the n tests are truly null, then

$$E[F(\tau)] = n_0 \cdot \Pr(p \leq \tau \,|\, \text{null hypothesis}) \simeq n_0 \cdot \tau \tag{A4.19}$$

Hence,

$$E[FDR(\tau)] = \frac{n_0 \cdot \tau}{S(\tau)} \tag{A4.20}$$

Setting $\tau = p(k)$, then $S(\tau) = k$, and Equation A4.20 becomes $n_0\, p(k)/k$, recovering Equation A4.16. Using the Storey-Tibshirani estimator (Equation A4.8b), an estimated value for the FDR using threshold value τ (and based on tuning parameter λ) becomes

$$\widehat{FDR}(\tau) = n_0 \cdot \frac{\tau}{S(\tau)} = \left(\frac{\text{Number of } p(i) \text{ values } > \lambda}{1-\lambda}\right) \cdot \left(\frac{\tau}{\text{Number of } p(i) \text{ values } \leq \tau}\right) \tag{A4.21}$$

Ideally, over a reasonable range of λ values, we expect this estimate 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.

Under a mixture model setting (e.g., Equation A4.9), the false discover rate given significance threshold τ is simply the fraction of all true positives declared significant to the fraction of all tests declared significant (i.e., those tests for which $p \leq \tau$). This can be estimated directly from the parameters of the mixture distribution,

$$\text{FDR}(\tau) = \frac{\pi_0\, \text{cdf}_U(\tau)}{\pi_0\, \text{cdf}_U(\tau) + (1-\pi_0)\, \text{cdf}_A(\tau)} = \frac{\pi_0\, \tau}{\pi_0 \tau + (1-\pi_0)\text{cdf}_A(\tau)}, \tag{A4.22}$$

where cdf denotes the cumulative distribution function,

$$\text{cdf}_U(x) = \int_0^x \phi_U(p)dp = x, \quad \text{cdf}_A(x) = \int_0^x \phi_A(p)dp$$

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 of tests. 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 *within the current experiment* whose p values are 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,

$$\widehat{q}\,[p(i)] = \min_{\tau \geq p(i)} \widehat{FDR}(\tau) \tag{A4.23}$$

Example A4.9: 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 breast cancer 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% (3.2/51) 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. 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 estimated the fraction π_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 *for this experiment* 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 A4.8, where the Benjamini-Hochberg estimator for FDR value was used (which differs from other FDR estimators by a constant, n_0/n). Notice that the smallest FDR occurs for hypothesis 6 (1.5%), and not for smaller p values. This reflects the tradeoff where increasing the threshold $p \leq \tau$ for significant results in declaring more tests as significant, so that the ratio $\tau/S(\tau)$ need not monotonically increase as τ increases. As example A4.8 shows, setting the threshold τ *above* the $p(i)$ value may actually result in a smaller q value, and hence Storey's definition.

Closing Caveats in using the FDR

While controlling the FDR is a very powerful approach for many multiple comparison problems, it is not a panacea. One concern is correlations among tests. As mentioned, in this case the null distribution of p values can significantly depart from a uniform, giving biased estimates of π_0 (and thus FDR). Further, recall that FDR control is accomplished by controlled the *expected* value of the FDR (or some closely related measure such as the PFP). The *variance* in the FDR across independent experiments can be considerable, especially when tests are correlated (Owen 2005; Leek and Storey 2011). One approach for treating these concerns is to use Leek and Storey's (2007, 2008) surrogate variable analysis to account for dependencies among the data before the actual p values for individual tests are obtained.

A second issue is a bit more subtle. Consider a standard QTL mapping experiment (LW Chapter 15) wherein a controlled cross is made between two lines (typically inbred) and one looks for marker-trait associations in the resulting F_2 (or other) progeny by scanning for linkage signals across a number of linked markers that span each chromosome. For each marker, the null hypothesis is no linkage to a QTL influencing the trait, while the alternative is that the marker is linked to a QTL. As noted by Chen and Storey (2006), the linkage signal from a QTL influences essentially all the markers on the chromosome on which it resides, and so *as a group* they all satisfy the same hypothesis. Either all are nulls (unlinked to a QTL) or failures of the null (linked to a QTL, albeit with differing degrees of a linkage signal). As such, an investigator can arbitrary obtain any level by simply adding or subtracting linked markers, and FDR control is not appropriate for this setting. To a much lesser extent, this same issue occurs in genome wide association studies among sets of extremely tightly linked SNPs. However, since the linkage signal in these cases is the persistence of linkage disequilibrium (LD) over large number of generations, any common signal is restricted to a set of very tightly linked markers, rather than an entire chromosome, and control of FDR among such clusters is 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. [A4]

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. [A4]

Benjamini, Y., and Hochberg, T. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 26: 60–83. [A4]

Chen, H., J. Chen, and J. D. Kalbfleisch. 2001. A modified likelihood ratio test for homogeneity in finite mixture models. *Stat. Meth.* 63: 19–29. [A4]

Chen, L., and J. D. Storey. 2006. Relaxed significance criteria for linkage analysis. *Genetics* 173: 2371–2381. [A4]

Dickersin, K., Y.-I. Min, and C. L. Meinert. 1992. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *J. Amer. Med. Assoc.* 267: 374–378. [A4]

Easterbrook, P. J., R. Gopalan, J. A. Berlin, and D. R. Matthews. 1991. Publication bias in clinical research. *The Lancet* 337: 867–872. [A4]

Efron, B. 2007. Correlation and large-scale simultaneous significance testing. *J. Amer. Stats Assoc.* 102: 93–103. [A4]

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. [A4]

Fisher, R. A. 1932. *Statistical methods for research workers.* Oliver and Boyd, Edinburgh. [A4]

Gadbury, G. L., G. P. Page, J. Edwards, T. Kayo, T. A. Prolla, R. Weindruch, P. A. Permana, J. D. Mountz, and D. B. Allison. 2004. Power and sample size estimation in high dimensional biology. *Stat. Methods Med. Res.* 13: 325–338. [A4]

Gadbury, G. L., Q. Xiang, L. Yang, S. Barnes, G. P. Page, and D. B. Allison. 2008. Evaluating statistical methods using plasmode data sets in the age of massive publich databases: An illustration using false discovery rates. *PLoS Genetics* 6: e1000098. [A4]

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. [A4]

Guedj, M., S. Robin, A. Celisse, and G. Nuel. 2009. Kerfdr: a semi-parametric kernel-based approach to local false discovery rate estimation. *BMC Bioniformatics* 10: 84. [A4]

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75: 800–802. [A4]

Holm, S. 1979. A simple sequential rejective multiple test procedure. *Scand. J. Statistics* 6: 65–70. [A4]

Hommel, G. 1988. A stagewise rejective multiple test procedure on a modified Bonferroni test. *Biometrika* 75: 383 – 386. [A4]

Hommel, G. 1989. A comparison of two modified Bonferonii procedures. *Biometrika* 76: 624-625. [A4]

Hu, X., G. L. Gadbury, Q. Xiang, and D. B. Allison. 2011. Illustrations on using the distribution of a p-value in high dimensional data analysis. *Adv. Appl. Stat. Sci.* 1: 191–213. [A4]

Leek, J. T., and J. D. Storey. 2007. Capturing heterogeneity in gene expression studies by surrogate variables. *PLoS Genetics* 3: e161. [A4]

Leek, J. T., and J. D. Storey. 2008. A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci.* 105: 18718–18723. [A4]

Leek, J. T., and J. D. Storey. 2011. The joint null criterion for multiple hypothesis tests. *Stat. Appl. Gen. Mol. Biol.* 10: Issue 1, Article 28. [A4]

Liptak, T. 1958. On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* 3: 171–197. [A4]

Manly, K. F., D. Nettleton, and J. T. G. Hwang. 2004. Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res.* 14: 997–1001. [A4]

McLachlan, G. J. 1987. On bootstapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Stat.* 36: 318–324. [A4]

Morton, N. E. 1955. Sequential tests for the detection of linkage. *American Journal of Human Genetics* 7: 277–318. [A4]

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. [A4]

Mosteller, F., and R. R. Such. 1954. Selected quantitative techniques. *In*, G. Lindzer (ed), *Handbook of social psychology*, Vol. 1, pp. 289–334. Addison-Wesley, Cambridge, Ma. [A4]

Nettleton, D., J. T. G. Hwand, R. A. Caldo, and R. P. Wise. 2006. Estimating the number of true null hypotheses from a histogram of p values. *J. Agr. Biol. Envir. Stats* 11: 337–356. [A4]

Owen, A. B. 2005. Variance of the number of false discoveries. *J. R. Statist. Soc.* B 67: 411–426. [A4]

Pearson, E. S. 1938. The probability transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* 30: 134–148. [A4]

Rice, W. H. 1990. A consensus combined p-value test and the family-wide significance of component tests. *Biometrics* 46: 303–308. [A4]

Robin, S., A. Bar-Hen, J. J. Daudin, and L. Pierre. 2007. A semi-parametic approach for mixture models: application to local false discovery rate estimation. *Comput. Statist. and Data Analysis* 51: 5483–5493. [A4]

Schweder, T., and E. Spjøtvoll. 1982. Plots of p-values to evaluate many tests simultaneously. *Biometrika* 69: 493–502. [A4]

Schork, N. 1992. Bootstrapping likelihood ratios in quantitative genetics. *In* R. Lepage and L. Billard (eds.), *Exploring the limits of the bootstrap*, pp. 389–393. Wiley, New York. [A4]

Shaffer, J. P. 1995. Multiple hypothesis testing. *Ann. Rev. Psychol.* 46: 561–584. [A4]

Simes, J. R. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 75–754. [A4]

Storey J.D. 2002. A direct approach to false discovery rates. *J. Royal Stat. Soc. Series B:* 64: 479–498. [A4]

Storey J.D. 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics* 31: 2013-2035. [A4]

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. *J. Royal Stat. Soc., Series B* 66: 187-205. [A4]

Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100: 9440–9445. [A4]

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. 1949. *The American solder, Vol. 1: Adjustment during army life.* Princeton University Press , Princeton. [A4]

Tippett, L. H. 1931. *The methods of statistics.* Williams and Norgate, London. [A4]

Whitlock, M. C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* 18: 1368–1373. [A4]