Statistics 210B, Spring 1998

Class Notes

P.B. Stark

stark@stat.berkeley.edu

www.stat.berkeley.edu/∼stark/index.html

April 24, 2012

Ninth Set of Notes

1 Multiplicity

For references, see J. Shaffer (1995) Multiple Hypothesis Testing, *Ann. Rev. Psychol., 46*, 561-584; J. Hsu (1996) *Multiple Comparisons: Theory and Methods*, Chapman and Hall, London.

It is often the case that one wishes to test not just one, but several or many hypotheses. For example, one might be evaluating a collection of drugs, and want to test the family of null hypotheses that each is not effective. Suppose one tests each of these null hypotheses at level α. This level is called the "per-comparison error rate" (PCER). Clearly, the chance of making at least one Type I error is at least α, and is typically larger. Let $\{H_j\}_{j=1}^m$ (m for multiplicity) be the family of null hypotheses to be tested, and let $H = \cap_j H_j$ be the "grand null hypothesis." If H is true, the expected number of rejections is αm. The "familywise error rate" (FWER) is the probability of one or more incorrect rejections:

$$\text{FWER} = \mathbf{P}\{ \text{ reject one or more true } H_j \}. \tag{1}$$

Strong control of the FWER at level α means that the probability of one or more incorrect rejections is at most α, whichever of the hypotheses $\{H_j\}$ happen to be true. *Weak control*

of the FWER at level α means that the probability of one or more incorrect rejections when the "grand null hypothesis" H is true is at most α:

$$\mathbf{P}_H\{ \text{ reject one or more } H_j \ \} \leq \alpha. \tag{2}$$

Typically, the FWER is much larger than the significance level at which the individual hypotheses are tested. This "multiplicity problem" is quite commonly ignored, which tends to lead to an overstatement of the significance of results (*i.e.*, the true significance level of the overall test is larger than the reported significance level, so more results are reported to be significant than should be). One situation in which the problem is evident is in the bias towards publishing only results that are statistically significant, so the many tests performed in search of a significant one are not reported. The rejection of a null hypothesis is sometimes called a "statistical discovery," and the fraction of rejected null hypotheses that are incorrectly rejected (that are in fact true) is called the "false discovery rate" (FDR).

We'll use the notation in Benjamini and Hochberg, 1995, Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing, *JRSS B*, *57*, 289-300.

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
False null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

The number m is the number of hypotheses tested; this is known. The number m_0 is the number of null hypotheses that are true; m_0 is not known. The random variable R is the total number of rejected null hypotheses; R is observable. The random variables U, V, T, and S are not observable.

If each hypothesis is tested individually at level α, then the PCER is $E(V/m) \leq \alpha$.

The same multiplicity issues arise in computing confidence intervals: If $\{I_j\}_{j=1}^m$ are individually level $1-\alpha$ confidence intervals for a set of parameters $\{\mu_j\}_{j=1}^m$, so that

$$\mathbf{P}\{I_j \ni \mu_j\} = 1 - \alpha, \quad j = 1, \ldots, m, \tag{3}$$

then the event

$$A = \cap_{j=1}^m \{I_j \ni \mu_j\} \tag{4}$$

typically has probability much less than $1 - \alpha$.

2 Controlling the FWER

2.1 Bonferroni Procedures

Bonferroni's inequality says that for any collection of events $\{A_j\}$, $\mathbf{P}\{\cup_j A_j\} \leq \sum_j \mathbf{P}A_j$. Thus the chance of one or more type I errors in an arbitrary collection of tests is at most the sum of their separate chances of type I errors. Thus if the hypotheses are tested individually at level α/m, the FWER is $\mathbf{P}\{V \geq 1\} \leq \alpha$.

Holm's Sequentially Rejective Bonferroni Method is based on the ordered p-values $p_1 \leq p_2 \leq \ldots \leq p_m$ of the m hypotheses tested; the hypotheses are assumed to be similarly ordered (the p-value of H_1 is p_1, *etc.*). Holm's test is

$$\text{Reject } H_i \text{ if } p_k \leq \alpha/(m-k+1) \text{ for all } k \leq i.$$
$$\text{Do not reject the other hypotheses.}$$

Theorem 1 *Holm's method controls the FWER at level α.*

Proof. Let m_0 be the number of true null hypotheses. If $m_0 = m$, there is an incorrect rejection only if $p_1 \leq \alpha/m$, which has probability α. If $m_0 = m-1$, there is an incorrect rejection if either H_1 is one of the true null hypotheses and $p_1 \leq \alpha/n$, or if H_1 is false, $p_1 \leq \alpha/m$, and $p_2 \leq \alpha/(m-1)$. Let p'_j be the jth smallest p-value among the m_0 true null hypotheses. There can only be an incorrect rejection if $p'_1 \leq \alpha/(m-1)$ (but that condition is not sufficient for an incorrect rejection). Thus the chance of an incorrect rejection is at most α. One can proceed similarly for $m_0 = m-2, \ldots, 1$, arguing that an incorrect rejection can only occur (but does not necessarily occur) if $p'_1 \leq \alpha/m_0$; in each case, the chance is at most α. Obviously, if $m_0 = 0$, there can be no incorrect rejection.

Holm's method is an example of a step-down procedure. The schematic of a step-down procedure is that one looks at the smallest p-value first. If that is larger than some threshold, no hypothesis is rejected. If not, the corresponding hypothesis is rejected, and one goes on to the second-smallest p-value. As soon as one reaches the point that the jth smallest p-value is larger than the jth threshold, no more hypotheses are rejected.

In a step-up procedure, one looks first at the largest p-value. If that is sufficiently small, all the hypotheses are rejected. If not, the corresponding hypothesis is not rejected, and one goes on to the second largest p-value. As soon as one reaches the point that the jth largest p-value is smaller than the jth threshold, all the remaining hypotheses are rejected.

2.2 Independent Test Statistics

Suppose we wish to test with FWER not exceeding α the family of hypotheses $\{H_i\}_{i=1}^m$ using independent test statistics $\{T_i\}_{i=1}^m$. Suppose we test each hypothesis at level β. Then the probability of one or more incorrect rejections (the FWER) is $1-(1-\beta)^m$. To have FWER equal to α requires

$$\begin{aligned} \alpha &= 1-(1-\beta)^m \\ (1-\alpha)^{1/m} &= 1-\beta \\ \beta &= 1-(1-\alpha)^{1/m}. \end{aligned} \tag{5}$$

Thus if we test the hypotheses individually at level $1-(1-\alpha)^{1/m}$, the FWER is at most α.

2.2.1 Simes' inequality.

See R.J. Simes (1986) An improved Bonferroni procedure for multiple tests of significance, *Biometrika, 73*, 751-754.

Suppose we are testing m null hypotheses $\{H_j\}$ using independent test statistics T_j. Let P_j be the jth smallest p-value among the m p-values. Simes' method is

reject the grand null hypothesis if for some j, $P_j \le j\alpha/m$.

Simes' method has FWER at most α.

Theorem 2 *(Simes, 1986). Let P_j be the jth order statistic of m iid $U(0,1)$ random variables. Then for $\alpha \in [0,1]$,*

$$A_m(\alpha) = \mathbf{P}\{P_j > j\alpha/m;\ j=1,\ldots,m\} = 1-\alpha. \tag{6}$$

Proof. The proof is by induction on m. Clearly, the statement is true for $m=1$. For $m>1$, $\{P_1/P_m, P_2/P_m, \ldots, P_{m-1}/P_m\}$ are distributed as the order statistics of $m-1$ iid $U(0,1)$ random variables, independent of P_m. Thus, for $p \ge \alpha$,

$$\begin{aligned} \mathbf{P}\{P_j > \frac{j\alpha}{m};\ j=1,\ldots,m-1 | P_m = p\} &= \mathbf{P}\{P_j/p > \frac{j\alpha(m-1)}{m(m-1)p};\ j=1,\ldots,m-1 | P_m = p\} \\ &= \mathbf{P}\{P_j > \frac{j\frac{(m-1)\alpha}{pm}}{m-1};\ j=1,\ldots,m-1 | P_m = p\} \\ &= A_{m-1}\left(\frac{(m-1)\alpha}{pm}\right). \end{aligned} \tag{7}$$

The distribution function of P_m is p^m, $p \in [0,1]$, so the density of P_m is mp^{m-1}. Suppose $A_{m-1}(\alpha) = 1-\alpha$, $\alpha \in [0,1]$.

$$\begin{aligned} A_m(\alpha) &= \int_\alpha^1 A_{m-1}\left(\frac{(m-1)\alpha}{pm}\right) mp^{m-1}dp \\ &= \int_\alpha^1 (1-\frac{(m-1)\alpha}{pm}) mp^{m-1}dp \\ &= p^m|_\alpha^1 - \alpha p^{m-1}\Big|_\alpha^1 \\ &= 1-\alpha^m-\alpha+\alpha^m \\ &= 1-\alpha. \end{aligned} \tag{8}$$

Note that $j\alpha/m$ is at least as large as $\alpha/(m-j+1)$, so the grand null is rejected more frequently using this test than using Holm's Bonferroni-based test.

Simes' result has recently been generalized to *positively regression dependent test statistics*. Two random variables X and Y are positively regression dependent if for $x_0 < x_1$, a random variable that has the conditional distribution of Y given $X = x_1$ is stochastically larger than that of one with the conditional distribution of Y given $X = x_0$ (Y tends to be larger when X is larger). Positively correlated normal random variables have positive regression dependence.

2.3 Chebychev's Other Inequality

Theorem 3 *(J. Hsu, 1996,* Multiple Comparisons: Theory and Methods*, Chapman and Hall, London. Theorem A.1.1.) Let X be an n-dimensional random variable. Suppose the functions $f, g : \mathbf{R}^n \to \mathbf{R}$ satisfy*

$$[f(x_2)-f(x_1)][g(x_2)-g(x_1)] \geq 0 \tag{9}$$

for all x_1, x_2 in the support of the distribution on X. Then, provided the expectations exist,

$$E[f(X)g(X)] \geq E[f(X)]E[g(x)]. \tag{10}$$

I.e., *$f(X)$ and $g(X)$ are positively correlated.*

Proof. Let X, X_1 and X_2 be iid. Then

$$\begin{aligned} 0 &\leq E\left[(f(X_2)-f(X_1))(g(X_2)-g(X_1))\right] \\ &= E\left[(f(X_2)g(X_2)+f(X_1)g(X_1)) - (f(X_1)g(X_2)+f(X_2)g(X_1))\right] \\ &= 2\left[E[f(X)g(X)] - E[f(X)]E[g(X)]\right]. \end{aligned} \tag{11}$$

Corollary 1 *(Kimball's inequality; Hsu, Corollary A.1.1.) Let V be a univariate random variable. If $\{g_j\}_{j=1}^m$ are bounded, nonnegative real functions, monotone in the same direction, then*

$$E\left[\Pi_{j=1}^m g_j(V)\right] \geq \Pi_{j=1}^m E g_j(V). \tag{12}$$

Proof. Use induction from two functions to m functions in the Theorem, taking $n = 1$.

2.4 Application to the one-way model

Suppose we are interested in the *one-way model*. We observe

$$X_{ia} = \mu_i + \epsilon_{ia}, \;\; i = 1, \ldots, m; \; a = 1, \ldots, n_i. \tag{13}$$

This is a model for making n_i observations of the response to treatment i for m different treatments, under the assumption that the response is a mean response plus a random effect. Assume that the errors ϵ_{ia} are iid $N(0, \sigma^2)$, σ^2 unknown. Let $\hat{\mu}_i = \bar{X}_i = \frac{1}{n_i}\sum_{a=1}^{n_i} X_{ia}$. Let $\nu = \sum_{i=1}^m (n_i - 1)$, and define

$$\hat{\sigma}^2 = \frac{1}{\nu}\sum_{i=1}^m \sum_{a=1}^{n_i} (X_{ia} - \bar{X}_i)^2. \tag{14}$$

The estimators $\{\hat{\mu}_i\}$ are independent normals with means $\{\mu_i\}$ and variances $\{\sigma^2/n_i\}$, independent of $\hat{\sigma}^2$, and $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$. For future use, define

$$\hat{\mu} = (\hat{\mu}_j)_{j=1}^m. \tag{15}$$

and

$$\hat{\sigma}_B^2 = \frac{1}{m}\sum_{i=1}^m n_i \left(\hat{\mu}_i - \frac{1}{m}\sum_{i=1}^m \hat{\mu}_i\right)^2. \tag{16}$$

Suppose we wish to find simultaneous confidence intervals for the set of parameters $\{\mu_i\}$. Define the Studentized test statistics

$$T_i = \frac{\hat{\mu}_i - \mu_i}{\hat{\sigma}/\sqrt{n_i}}, \; i = 1, \ldots, m. \tag{17}$$

These test statistics are dependent, because of the common divisor $\hat{\sigma}$. If they were not, the intervals

$$[\hat{\mu}_i - \hat{\sigma} t_{1-(1-(1-\alpha/2)^{1/m})/2}, \hat{\mu}_i + \hat{\sigma} t_{1-(1-(1-\alpha/2)^{1/m})/2}], \; i = 1, \ldots, m \tag{18}$$

would be exact $1 - \alpha$ simultaneous confidence intervals for $\{\mu_i\}$.

Kimball's inequality lets one show that these intervals are in fact conservative as a result of the dependence on $\hat{\sigma}$. Let A_i be the event that the ith interval covers. Consider the function $g_i(\hat{\sigma}) = \mathbf{P}\{A_i|\hat{\sigma}\}$. These functions all increase monotonically with $\hat{\sigma}$. Recall that $\{\hat{\mu}_i\}$ are independent of each other and of $\hat{\sigma}$. Thus

$$\begin{aligned}
\mathbf{P}\left\{\cap_{i=1}^m A_i\right\} &= E1_{\cap_{i=1}^m A_i} \\
&= E\Pi_{i=1}^m 1_{A_i} \\
&= E_{\hat{\sigma}} E(\Pi_{i=1}^m 1_{A_i}|\hat{\sigma}) \\
&= E_{\hat{\sigma}} \Pi_{i=1}^m \mathbf{P}\{A_i|\hat{\sigma}\} \\
&\geq \Pi_{i=1}^m E_{\hat{\sigma}} \mathbf{P}\{A_i|\hat{\sigma}\} \\
&= \Pi_{i=1}^m \mathbf{P}\{A_i\} \\
&= 1-\alpha, \qquad (19)
\end{aligned}$$

where Kimball's inequality was used in the penultimate step.

2.5 Comparisons and Constrasts

We specialize to the case that we are interested in a collection of m parameters $\{\mu_i\}_{i=1}^m$. Let $\mu = (\mu_i)_{i=1}^m$. The hypotheses we wish to test involve comparing the parameters or linear combinations of the parameters. For example, we might be interested in the family of hypotheses $\{H_{ij} : \mu_i = \mu_j\}_{i=1,\ldots,m-1;j=i+1,\ldots,m}$ (all pairwise comparisons). For $\mathcal{I}$ a subset of $\{1,\ldots,m\}$, let $H_{\mathcal{I}}$ denote the hypothesis that all $\{\mu_i\}_{i\in\mathcal{I}}$ are equal (perhaps a better notation would be that $\#\{\mu_i\}_{i\in\mathcal{I}} = 1$). We might be interested in the family of hypotheses $\{H_{\mathcal{I}}\}_{\mathcal{I}\in\mathcal{I}}$, where $\mathcal{I}$ is a collection of subsets of $\{1,\ldots,m\}$.

A *contrast* is a linear combination $\sum_{i=1}^m c_i\mu_i = c\cdot\mu$, with the restriction that $\sum_{i=1}^m c_i = c\cdot\mathbf{1} = 0$. A pairwise comparison is a constrast with $c_i = 1$ for some i, $c_j = -1$ for some $j \neq i$, and all other components of c equal to zero.

We are going to assume that we are in a one-way model with and equal number N of observations of each of the m treatments. We again assume that the observational errors are iid $N(0,\sigma^2)$, with σ^2 unknown. The rest of the notation is as in section 2.4.

The "cost" in terms of reduced power tends to increase with the number of hypotheses tested; if one is not interested in testing all possible contrasts, one can have more power testing the limited set. Some major divisions of families of hypotheses tested in the one-way

model include, in decreasing order of complexity, ACC (all contrasts comparison), MCA (all pairwise comparisons), MCB (multiple comparisons with the [sample] best), and MCC (multiple comparisons with control). MCC involves the fewest comparisons: $m-1$ sample values are compared with the mth, which is the control. In MCB, there are also only $m-1$ comparisons, but the measured effect $\hat{\mu}_i$ that the other $m-1$ are compared with is that one observed to be best; under the grand null, that is equally likely to be any of the $\hat{\mu}_i$. In MCA, there are $(m^2-m)/2$ hypotheses tested, and in ACC, an infinite number are tested.

2.5.1 The Scheffé Method

The Scheffé method controls the FWER for all possible contrasts (ACC). The "grand null" in this case is that all the μ_i are equal, so all contrasts are zero.

Recall that if Y has a chi-square distribution with k degrees of freedom and Y' has a chi-square distribution with ℓ degrees of freedom, and Y and Y' are independent, then

$$\frac{Y/k}{Y'/\ell} \tag{20}$$

has an F distribution with k and ℓ degrees of freedom, denoted $F_{k,\ell}$. Let $F_{k,\ell,\alpha}$ denote the α critical value of $F_{k,\ell}$. It is a standard result in the analysis of variance that under the one-way normal model, $(m-1)\hat{\sigma}_B^2/\sigma^2 \sim \chi^2_{m-1}$ and $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$ are independent, so

$$\frac{\hat{\sigma}_B^2}{\hat{\sigma}^2} \sim F_{m-1,\nu}. \tag{21}$$

The variables $\{\sqrt{n_i}(\hat{\mu}_i - \mu_i)/\sigma\}$ are iid $N(0,1)$, and $\sigma^{-2}\sum_{i=1}^m n_i(\hat{\mu}_i-\mu_i)^2$ has a chi-square distribution with m degrees of freedom, and is independent of $\hat{\sigma}^2$, so whatever be $\{\mu_i\}_{i=1}^m$,

$$\mathbf{P}\left\{\frac{\sum_{i=1}^m n_i|\hat{\mu}_i-\mu_i|^2}{m\hat{\sigma}^2} \le F_{m,\nu,\alpha}\right\} = 1-\alpha. \tag{22}$$

Equivalently,

$$\mathbf{P}\{\sum_{i=1}^m n_i|\hat{\mu}_i-\mu_i|^2 \le m\hat{\sigma}^2 F_{m,\nu,\alpha}\} = 1-\alpha. \tag{23}$$

In the case all $n_i = N$, this becomes

$$\mathbf{P}\{\|\hat{\mu}-\mu\|^2 \le \frac{m}{N}\hat{\sigma}^2 F_{m,\nu,\alpha}\} = 1-\alpha. \tag{24}$$

That is, the chance is at least $1-\alpha$ that $\hat{\mu} \in \mathbf{R}^m$ is in a ball centered at μ of radius

$$r_\alpha = \frac{m}{N}\hat{\sigma}\sqrt{\frac{mF_{m,\nu,\alpha}}{N}}. \tag{25}$$

The unit ball in $\mathbf{R}^m$ can be characterized as

$$\{\beta \in \mathbf{R}^m : |c \cdot \beta| \le \|c\|\}, \tag{26}$$

so

$$\mathbf{P}\{|c \cdot \hat{\mu} - c \cdot \mu| \le \|c\| r_\alpha \; \forall c \in \mathbf{R}^m\} = 1 - \alpha. \tag{27}$$

This gives simultaneous confidence intervals for $c \cdot \mu$ (whether or not c is a "contrast") as

$$\mathcal{I}_c = [c \cdot \hat{\mu} - \|c\| r_\alpha, c \cdot \hat{\mu} + \|c\| r_\alpha]. \tag{28}$$

For testing contrasts, one rejects the hypothesis that $c \cdot \mu = 0$ if $|c \cdot \hat{\mu}| > \|c\| r_\alpha$, and one rejects the grand null hypothesis if $\|\hat{\mu}\| \ge r_\alpha$. Any number of contrasts can be tested this way, with FWER strongly controlled at level α.

Note that if one uses Scheffé's method to produce confidence intervals only for the effects $\{\mu_i\}$, it is unnecessarily conservative: it amounts to projecting a ball onto the coordinate axes, which is equivalent to taking the corresponding hyperrectangle as the confidence set for μ. That hyperrectangle strictly contains the ball, so it has higher coverage probability than the ball. If we were interested only in simultaneous confidence intervals for $\{\mu_i\}$, we could get shorter confidence intervals by starting with a hyperrectangular confidence region for μ (with faces aligned with the axes), and projecting *that* set. This is more or less what Tukey's maximum modulus method does.

2.5.2 Tukey's Maximum Modulus Method

Tukey's method was originally introduced for all pairwise comparisons, but can be modified for ACC. Again, let's take $n_i = N$. Define $c^*(\alpha)$ to satisfy

$$\mathbf{P}\left\{\frac{|\hat{\mu}_i - \hat{\mu}_j - (\mu_i - \mu_j)|}{\hat{\sigma}\sqrt{2/N}} \le c^*(\alpha) \forall j < i\right\} = 1 - \alpha. \tag{29}$$

Values of $c^*(\alpha)$ can be found by numerical integration. Then

$$\mathcal{I}_{ij} = [\hat{\mu}_i - \hat{\mu}_j - c^*(\alpha)\hat{\sigma}\sqrt{2/N}, \hat{\mu}_i - \hat{\mu}_j - c^*(\alpha)\hat{\sigma}\sqrt{2/N}], \; j < i \tag{30}$$

are simultaneous level $1-\alpha$ confidence intervals for the $(m^2 - m)$ pairwise difference $\mu_i - \mu_j$, $j < i$. By construction, the tests

$$\text{reject } H_{ij} : \mu_i = \mu_j \text{ if } |\hat{\mu}_i - \hat{\mu}_j| > c^*(\alpha)\hat{\sigma}\sqrt{2/N}$$

control the FWER for all pairwise comparisons at level α.

3 The False Discovery Rate

See Benjamini and Hochberg (1995) for the development of this idea. It's quite new, and quite promising in a variety of settings, including nonparametric function estimation, where the hypotheses tested are that individual coefficients of the unknown function in some basis expansion are zero.

Let $Q = V/(V + S)$ be the fraction of rejected hypotheses that are rejected incorrectly. Define $Q = 0$ if $V+S = 0$ (no hypothesis could have been rejected incorrectly if no hypothesis was rejected).

The FDR is $Q_e = EQ = E\{V/(V + S)\}$, the expected fraction of rejected hypotheses that are incorrectly rejected. Note that

1. If all null hypotheses are true, the FDR is the same as the FWER: $\mathbf{P}\{V \geq 1|m_0 = m\} = E(Q)$. Controlling the FDR thus controls the FWER in a weak sense.

2. When the number m_0 of true null hypotheses is less than the total number m of hypotheses, FDR $\leq$ FWER. Thus controlling the FWER controls the FDR.

3.1 A procedure that controls the FDR

Let $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$ be the ordered p-values of the m hypotheses. Let $H_{(i)}$ be the hypothesis corresponding to the p-value $P_{(i)}$. Define

$$K(q^*) \equiv \max\{i : P_{(i)} \leq \frac{i}{m} q^*\}. \tag{31}$$

Consider the procedure "reject all $H_{(i)} : i \leq K(q^*)$."

Theorem 4 *Benjamini and Hochberg, 1995, Theorem 1. If the test statistics are independent, then for any configuration of false null hypotheses, this procedure controls the FDR at level q^*.*

The proof relies on the following lemma:

Lemma 1 *Benjamini and Hochberg, 1995, Lemma. Let the number of true null hypotheses be m_0, $0 \leq m_0 \leq m$. Order the hypotheses such that the first m_0 are the true ones. Let*

$m_1 = m - m_0$ be the number of false null hypotheses. If the test statistics of the true null hypotheses are independent, for the procedure just given,

$$E(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq \frac{m_0}{m} q^*. \tag{32}$$

Proof of Lemma. Benjamini and Hochberg prove the lemma by induction. Suppose $m = 1$. Then the procedure rejects H_1 if $P_1 \leq q^*$. If $m_0 = 0$, no incorrect rejection can occur, so

$$E(Q|P_1 = p_1) = 0 \leq \frac{0}{1} q^*. \tag{33}$$

If $m_0 = 1$, an incorrect rejection occurs if $P_1 \leq q^*$. There is no P_{m_0+1}, so

$$E(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) = E(Q) = \mathbf{P}\{P_1 \leq q^*\} = q^* \leq \frac{1}{1} q^*. \tag{34}$$

Suppose that the lemma is true for all $m' \leq m$; we shall show that it is then true for $m' = m + 1$. If $m_0 = 0$, the null hypotheses are all false, so Q is identically zero, and the conditional expectation of Q, and so

$$E(Q|P_1 = p_1, \ldots, P_m = p_m) = 0 \leq \frac{m_0}{m+1} q^*. \tag{35}$$

Suppose $m_0 > 0$. Let P'_i, $i = 1, \ldots, m_0$ be the p-values corresponding to the true null hypotheses. Let $P'_{(m_0)}$ be the largest of the P'_i. Note that $\{P'_i\}_{i=1}^{m_0}$ are iid $U[0,1]$, so the density of $P'_{(m_0)}$ is $f(u) = m_0 u^{m_0 - 1}$. Let $\{p_j\}_{j=1}^{m_1}$ be the p-values of the false null hypotheses, so that $p_1 \leq p_2 \leq \cdots \leq p_{m_1}$. Define

$$j_0 \equiv \max\{j : p_j \leq \frac{m_0 + j}{m+1} q^*\}, \tag{36}$$

and

$$p_0 = \frac{m_0 + j_0}{m+1} q^*. \tag{37}$$

Note that $p_{j_0} \leq p_0$. Calculate the expectation, conditioning on the value of $P'_{(m_0)}$:

$$\begin{aligned} E(Q|P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) &= \int_0^{p_0} E(Q|P'_{(m_0)} = u, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f(u) du + \\ &+ \int_{p_0}^1 E(Q|P'_{(m_0)} = u, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f(u) du \end{aligned} \tag{38}$$

In the first integral, $u \leq p_0$, so all the null hypotheses are rejected, and $Q = \frac{m_0}{m_0 + j_0}$. Recall that $p_0 = \frac{m_0 + j_0}{m+1} q^*$. Thus

$$\int_0^{p_0} E(Q|P'_{(m_0)} = u, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f(u) du = \int_0^{p_0} E(\frac{m_0}{m_0 + j_0}|P'_{(m_0)} = u, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f$$

$$
\begin{aligned}
&= \int_0^{p_0} \frac{m_0}{m_0+j_0} m_0 u^{m_0-1} du \\
&= \frac{m_0}{m_0+j_0} p_0^{m_0} \\
&= \frac{m_0}{m_0+j_0} \frac{m_0+j_0}{m+1} q^* p_0^{m_0-1} \\
&= \frac{m_0}{m+1} q^* p_0^{m_0-1}.
\end{aligned}
$$

Consider the second integral. On the domain of the second integral, $P'_{(m_0)} = u \geq p_0 \geq p_{j_0}$. Here, at least the true null hypothesis with the largest p-value will not be rejected, and possibly others as well. Suppose $j > j_0$ so that $p_{j+1} \geq p_j \geq p_{j_0}$. Recall that $p_{j_0} \leq p_0$. Break the domain of integration into the intervals $p_j \leq u \leq p_{j+1}$, $j = j_0+1, \ldots, m_1 - 1$, together with $p_0 \leq u \leq p_{j_0+1}$ and $p_{m_1} \leq u \leq 1$. Because $u, p_{j+1}, \ldots, p_{m_1}$ are all greater than the threshold value p_0, their values cannot result in any hypothesis being rejected.

Let $\{H_{(i)}\}_{i=1}^m$ denote the entire set of m hypotheses, ordered by their p-values. In the second integral, the p-values of the true null hypotheses are all no larger than u (by definition of u—it's the largest p-value among the true null hypotheses). Recall that the rejection procedure is to reject all hypotheses with smaller p-values than p_0, so the rejection of $H_{(i)}$ implies that there must be some k, $i \leq k \leq m_0 + j - 1$ for which

$$
p_{(k)} \leq \frac{k}{m+1} q^*. \tag{40}
$$

This is equivalent to

$$
\frac{p_{(k)}}{u} \leq \frac{k}{m_0+j-1} \frac{m_0+j-1}{(m+1)u} q^*. \tag{41}
$$

The proof is now similar to that of Simes' inequality: conditional on $P'_{(m_0)} = u$, $\{P'_i/u\}_{i<m_0}$ are iid $U(0,1)$ random variables; $\{p_i/u\}_{i=1}^j$ are some numbers between 0 and 1 corresponding to false null hypotheses. We are testing $m_0 + j - 1 = m' < m$ hypotheses using a different value of q^*, namely $\frac{m_0+j-1}{(m+1)p} q^*$. The induction hypothesis gives This is like testing the $m' = m_0 + j - 1 \leq m$ using the threshold $\frac{m_0+j-1}{(m+1)p} q^*$. Because $m' \leq m$, we can apply the induction hypothesis:

$$
E(Q|P'_{(m_0)} = u, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) \leq \frac{m_0 - 1}{(m+1)u} q^*. \tag{42}
$$

This bound does not depend on p_j or p_{j+1}, so

$$
\begin{aligned}
\int_{p_0}^1 E(Q|P'_{(m_0)} = u, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}) f_{P_{(m_0)}}(u) du &\leq \int_{p_0}^1 \frac{m_0 - 1}{(m+1)u} q^* m_0 u^{m_0-1} du \\
&= \frac{m_0}{m+1} q^* \int_{p_0}^1 (m_0 - 1) u^{m_0-2} du
\end{aligned}
$$

$$= \frac{m_0}{m+1} q^* (1 - p_0^{m_0 - 1}). \qquad (43)$$

Adding this to the bound on the first integral proves the Lemma.

Proof of Theorem 4. Whatever be the joint distribution of $P_{m_0+1}, \cdots, P_m$ corresponding to the false null hypotheses, integrating the inequality in the Lemma gives

$$E(Q) = E(E(Q|P_{m_0+1}, \cdots, P_m) \leq \frac{m_0}{m} q^* \leq q^*. \qquad (44)$$

The FDR-controlling procedure is equivalent to picking α *a posteriori* to maximize the number $r(\alpha)$ of rejections at that level, subject to the constraint

$$\alpha m / r(\alpha) \leq q^*. \qquad (45)$$

That is, we reject as many hypotheses as possible, subject to the constraint that the expected number of incorrect rejections is at most the FDR times the number of hypotheses actually rejected. The expected number of incorrect rejections is $E(V) \leq \alpha m$, so $Q_e = EQ \leq \alpha m / r(\alpha) \leq q^*$.

One complaint about this FDR-controlling procedure (see the review article by Shaffer) is that because Q is defined to be zero when no rejection occurs, the conditional FDR given that some rejection does occur exceeds q^*.