

Department of Statistics
University of Central Florida

Technical Report TR-2007-01
25APR2007
Revised 25NOV2007

Controlling the Number of False Positives Using the Benjamini- Hochberg FDR Procedure
Paul N. Somerville
University of Central Florida

Controlling the Number of False Positives Using the Benjamini- Hochberg FDR Procedure
Paul N. Somerville
University of Central Florida

Version 26NOV2007

Abstract

For very large numbers of hypotheses, using the traditional family-wise error rate (FWER) can result in very low power for testing single hypotheses. Benjamini and Hochberg (1955) proposed a powerful multiple step procedure which controls FDR, the "False Discovery Rate". The procedure can result in a large number of false positives. Van der Laan, Dudoit and Pollard (2004) proposed controlling a generalized family-wise error rate k-FWER (also called gFWER(k)), defined as the probability of at least (k+1) Type I errors (k=0 for the usual FWER). Lehmann and Romano (2005) suggested new and simple methods of controlling k-FWER and the proportion of false positives (PFP) (also called False Discovery Proportion FDP). Somerville and Hemmelmann (2008) proposed controlling k-FWER by limiting the number of steps in step-up or step-down procedures. In this paper the procedure is applied to the Benjamini-Hochberg FDR procedure. Formulas are developed and Fortran 95 programs have been written. Tables are presented giving the maximum number of steps in the Benjamini-Hochberg procedure which will assure that

$$P[U \leq k] \geq 1 - \alpha$$

for various values of k and α , where U is the number of false positives.

Key words: false positives, controlling, Benjamini-Hochberg procedure, tables

1. INTRODUCTION

In multiple hypotheses testing, it is challenging to adequately control the rejection of true hypotheses while still maintaining reasonable power to reject false hypotheses. A standard procedure has been to control the family-wise error rate (FWER). The family-wise error rate is defined as the probability of rejecting at least one true hypothesis.

For very large numbers of hypotheses, using FWER can result in very low power for testing single hypotheses. Recently, powerful multiple step FDR procedures have been proposed which control the "False Discovery Rate". The false discovery rate is the expected proportion of the rejected hypotheses which are true (defined to be zero when no hypotheses have been rejected).

The first FDR procedure was proposed by Benjamini and Hochberg (1995). Although FDR procedures control the expected proportion of Type I errors, large numbers of false positives can result. Recently van der Laan, Dudoit and Pollard (2004) proposed controlling a generalized family-wise error rate k-FWER (also called gFWER(k)), defined as the probability of at least (k+1) Type I errors (k=0 for the usual FWER).

Lehmann and Romano (2005) suggested new methods of controlling gFWER(k) (called by them k-FWER) and PFP (called by them FDP - False Discovery Proportion). Both single-step and step-down procedures were derived which control the k-FWER. The procedures make no assumptions concerning the dependence structure or the p-values of the individual tests. The step-down procedure is simple to apply, and cannot be improved without violation of control of the k-FWER. Lehmann and Romano (2005) also proposed two methods for controlling the PFP (Proportion of False Positives). The first holds under "mild conditions on the dependence structure of p-values" while the second requires no dependence assumptions. They make no assumptions concerning the p-values of the individual tests.

Somerville and Hemmelmann (2008) proposed controlling k-FWER by limiting the number of steps in step-up or step-down procedures. In this paper the procedure is applied to the Benjamini-Hochberg FDR procedure. Formulas are developed and tables are presented.

2. STEP-UP AND STEP-DOWN PROCEDURES

Let t_1, t_2, \dots, t_m be test statistics corresponding to the null hypotheses H_1, H_2, \dots, H_m . Denote by T_i the random variable associated with t_i . Let $T_{(1)} \leq T_{(2)} \leq \dots \leq T_{(m)}$ be the ordered values for the test statistics and denote the corresponding hypotheses by $H_{(1)}, H_{(2)}, \dots, H_{(m)}$. Let d_i be constants such that $d_1 \leq d_2 \leq \dots \leq d_m$.

Step-down test procedures may be described as follows. If $T_{(m)} \geq d_m$, reject $H_{(m)}$ and continue for $i=m-1, m-2, \dots$, comparing $T_{(i)}$ with d_i , and rejecting $H_{(i)}$ if $T_{(i)} \geq d_i$, continuing until for the first time $T_{(i)} < d_i$. If $T_{(m)} < d_m$, reject no hypotheses. If $T_{(i)} > d_i$, for all values of i , reject all the hypotheses.

For step-up procedures $T_{(i)}$ is compared with d_i , beginning with $i = 1$, continuing for $i = 2, 3, \dots$ until, for

the first time, $T_{(i)} \geq d_i$. $H_{(i)}$, $H_{(i+1)}$, ..., $H_{(m)}$ are then rejected. If $T_{(i)} < d_i$, for all values of i , reject no hypotheses.

The procedures may also be formulated in terms of p-values and corresponding critical p-values.

3. REDUCED STEP PROCEDURES

An s-step, step-down procedure may be defined as follows. Compare $T_{(i)}$ with d_i for $i = m, m-1, \dots, m-s+1$ until for the first time $T_{(i)} < d_i$, in which case reject $H_{(m)}, \dots, H_{(i+1)}$. If the first time occurs for $i = m-s+1$, then also reject all hypotheses for which $T_j \geq d_{m-s+1}$. If $T_{(m)} < d_m$, reject no hypotheses.

For the s-step, step-up procedure, compare $T_{(i)}$ with d_i , for $i = m-s+1, m-s+2, \dots, m$ until for the first time $T_{(i)} \geq d_i$, in which case reject $H_{(i)}, H_{(i+1)}, \dots, H_{(m)}$. If $T_{(m-s+1)} \geq d_{m-s+1}$, reject all hypotheses for which $T_i \geq d_{m-s+1}$. If $T_{(m-s+1)} < d_{m-s+1}$, reject no hypotheses.

An s step procedure is equivalent to an m step procedure where the m-s smallest critical values are replaced with the value d_{m-s+1} . Thus, only the largest s critical values are used in the comparisons. The critical value d_{m-s+1} is the Minimum Critical Value (MCV). The concept of an MCV was introduced by Somerville (2004).

4. CONTROLLING k-FWER USING THE BENJAMINI-HOCHBERG PROCEDURE

Assume the m random variables T_i are independent and identically distributed. Then, by definition of the BH procedure,

$$\begin{aligned} P[d_i < T_i \leq d_{i+1}] &= e = q/m \\ P[T_i \geq d_m] &= e \\ P[T_i < d_1] &= 1 - q = 1 - me \end{aligned}$$

where $q (= \alpha)$ is the False Discovery Rate.

The following are also true:

$$\begin{aligned} P[T_i \geq d_j] &= (m-j+1)e \\ P[T_i < d_j] &= 1 - (m-j+1)e \\ P[T_i \geq \text{MCV}] &= se. \end{aligned}$$

Our object is to find the maximum number of steps s, using the BH procedure, for which

$$P[U \leq k] \geq 1 - \gamma \quad \text{given } k \text{ and } \gamma$$

where U is the random variable corresponding to the number of false discoveries.

Suppose H_1, H_2, \dots, H_f are the f false hypotheses with the means of the corresponding random variables T_i equal to $\mu_1, \mu_2, \dots, \mu_f$. Assume further that the random variables T_i have a distribution such that

$$P[T_i \geq a \mid \mu_i = \mu] = P[T_i \geq (a + \delta) \mid \mu_i = (\mu + \delta)] \quad \text{where } \mu, a \text{ and } \delta \text{ are arbitrary constants.}$$

Then, by theorem 6.1 of Somerville and Hemmelmann (2008), $P[U \leq k]$ is minimized when the means of the random variables corresponding to the false hypotheses increase without limit.

Assuming the least favorable locations of the means of the random variables corresponding to the false hypotheses, the false hypotheses will be rejected with probability 1. Then A_{f+i} , the probability that exactly i true hypotheses will be rejected, in addition to the f false hypotheses is given by

$$A_{f+i} = P[T_{(1)} < d_1, \dots, T_{(m-f-i)} < d_{m-f-i}; T_{(m-f-i+1)} \geq d_{m-f-i+1}].$$

Let s be the number of steps. Define j as follows:

$$s = f+i+j, \text{ or } s-j = f+i.$$

Then $\text{MCV} = d_{m-s+1} = d_{m-f-i-j+1}$.

If $j \leq 0$, $\text{MCV} \geq d_{m-f-i+1}$, and

$$\begin{aligned} A_{f+i} &= P[T_{(1)} < \text{MCV}, \dots, T_{(m-f-i)} < \text{MCV}; T_{(m-s+1)} \geq \text{MCV}] \\ &= R(m-f; i, 0) (1-se)^{m-f-i} (se)^i, \text{ where} \end{aligned}$$

$R(A; a_1, a_2) = R(A; a_1, a_2, a_3)$ is the number of partitions of A objects into three distinct categories containing a_1, a_2, a_3 respectively in the three categories.

If $j \geq 1$,

$$A_{f+i} = P[T_{(1)} < \text{MCV}, \dots, T_{(m-s+1)} < \text{MCV}, T_{(m-s+2)} < d_{m-s+2}, \dots, T_{(m-s+j)} < d_{m-s+j}; T_{(m-s+j+1)} \geq d_{m-s+j+1}]$$

Define three categories for the m test statistics:

- Category 1: $T_i < \text{MCV}$
- Category 2: $\text{MCV} \leq T_i < d_{m-s+j}$
- Category 3: $d_{m-s+j+1} \leq T_i$.

Note that: category 1 consists of $m-s+1$ intervals (blocks) $(-\infty, d_1), \dots, [d_{m-s}, d_{m-s+1})$; category 2 consists of the next $j-1$ intervals (blocks) $[d_{m-s+1}, d_{m-s+2}), \dots, [d_{m-s+j-1}, d_{m-s+j})$; and category 3 as consists of the last $s-j$ intervals (blocks) $[d_{m-s+j+1}, d_{m-s+j+2}), \dots, [d_m, \infty)$. Note that there are no test statistics in $[d_{m-s+j}, d_{m-s+j+1})$.

Using the formula for A_{f+i} , and the BH procedure, for exactly i true hypotheses to be rejected, at least i test statistics must be in each of the first i consecutive intervals for all $i \leq m-s+j$, with exactly $m-s+j$ test statistics in the first $m-s+j$ (category 1 plus category 2). The number of test statistics in category 1 will range from $m-s+1$ to $m-s+j$ and the number in category 2 will range from $j-1$ to 0.

If the number of test statistics in category 1 is $m-s+t$, then there will be $j-t$ in category 2 and the number of partitions of the $m-f$ or $(m-s+i+j)$ true hypotheses into the three categories is $R(m-f; m-s+t, j-t, i)$.

Let $C(j,t)$ be the number of ways for which the requirements of the BH procedure can be met for a partition for which there are exactly $m-s+t$ test statistics in category 1, $(j-t)$ in category 2) and i in category 3. Then

$$A_{f+i} = \sum_{t=1}^j R(m-s+i+j; m-s+t, j-t, i) * C(j,t) * (1-se)^{m-s+t} * e^{j-t} * (s-j)^i * e^i,$$

or

$$A_{f+i} = \sum_{t=1}^j R(m-f; m-f-i-j+t, j-t, i) * (1-se)^{m-f-i-j+t} * C(j,t) * e^{j-t} * (s-j)^i * e^i,$$

Let a_1, \dots, a_t be t test statistics in category 1 and a_{t+1}, \dots, a_j be $j-t$ test statistics in category 2. Let $(-\infty, d_{m-s+1})$ in category 1 be the first block and let $[d_{m-s+1}, d_{m-s+2}), [d_{m-s+2}, d_{m-s+3}), \dots, [d_{m-s+j-1}, d_{m-s+j})$ be the next $j-1$ blocks. Then, obtaining the value of $C(j,t)$ is equivalent to finding the number of partitions of the $j-t$ test statistics into the last $j-1$ blocks such that the number of elements in the first i blocks is at least i . Using the theorem in the appendix, $C(j,k) = k * j^{k-1}$.

5. FINDING THE MAXIMUM NUMBER OF STEPS FOR THE CONTROL OF $P[U \leq k]$

Applying the theorem of Somerville and Hemmelmann (2008), $P[U \leq k] = \sum_{i=0}^k A_{f+i}$ is minimized when the means of all of the test statistics corresponding to the false hypotheses increase without limit. A Fortran95 program BHstep has been written which finds the largest numbers of steps s for the BH procedure such that $P[U \leq k] \geq \gamma$ for arbitrary values of m, q, U and γ . The program starts with $s=1$, finding the value of f , the number of false hypotheses, for which $P[U \leq k]$ is smallest. It then increments s until $P[U \leq k]$ no longer equals or exceeds γ . The Fortran program will be submitted for publication.

6. SOME USEFUL TABLES AND GRAPHS

Figures 6.1, 6.2 and 6.3 show the maximum number of steps that result in $P[U \leq k]$ greater than or equal to .90, .95 and .99 respectively when $q = .05$ for various values of m ranging from 50 to 10^7 .

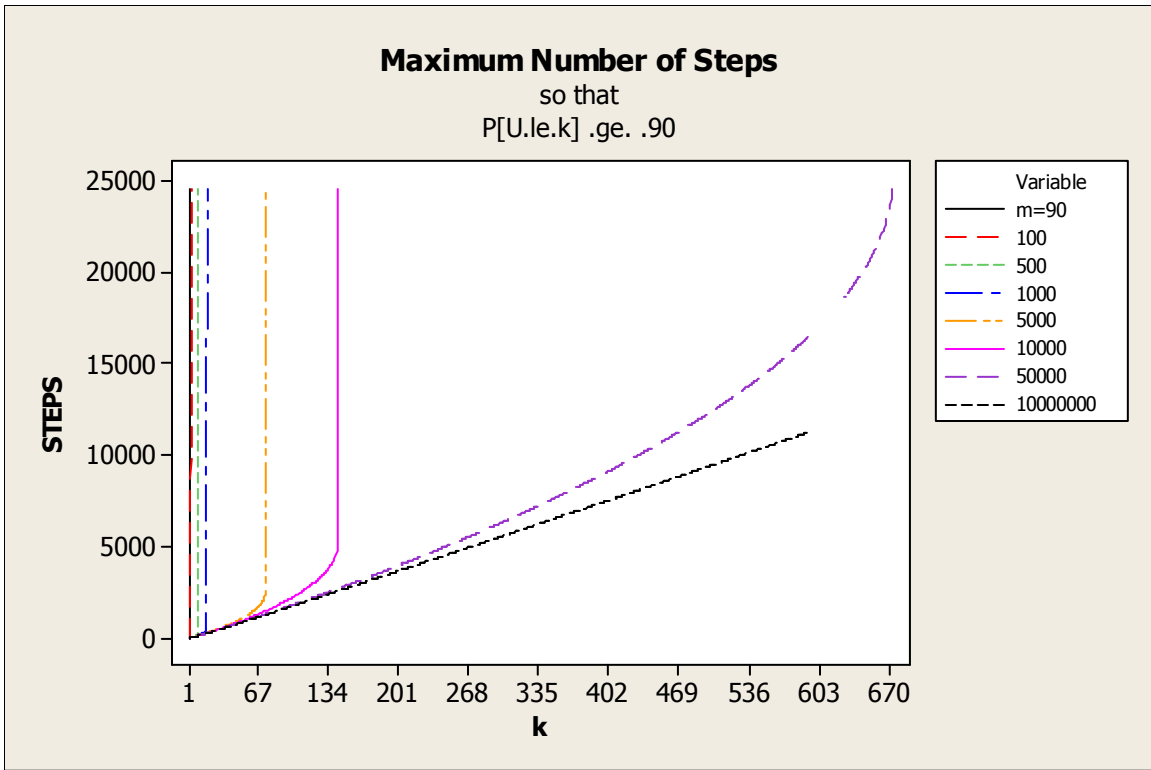


Figure 6.1

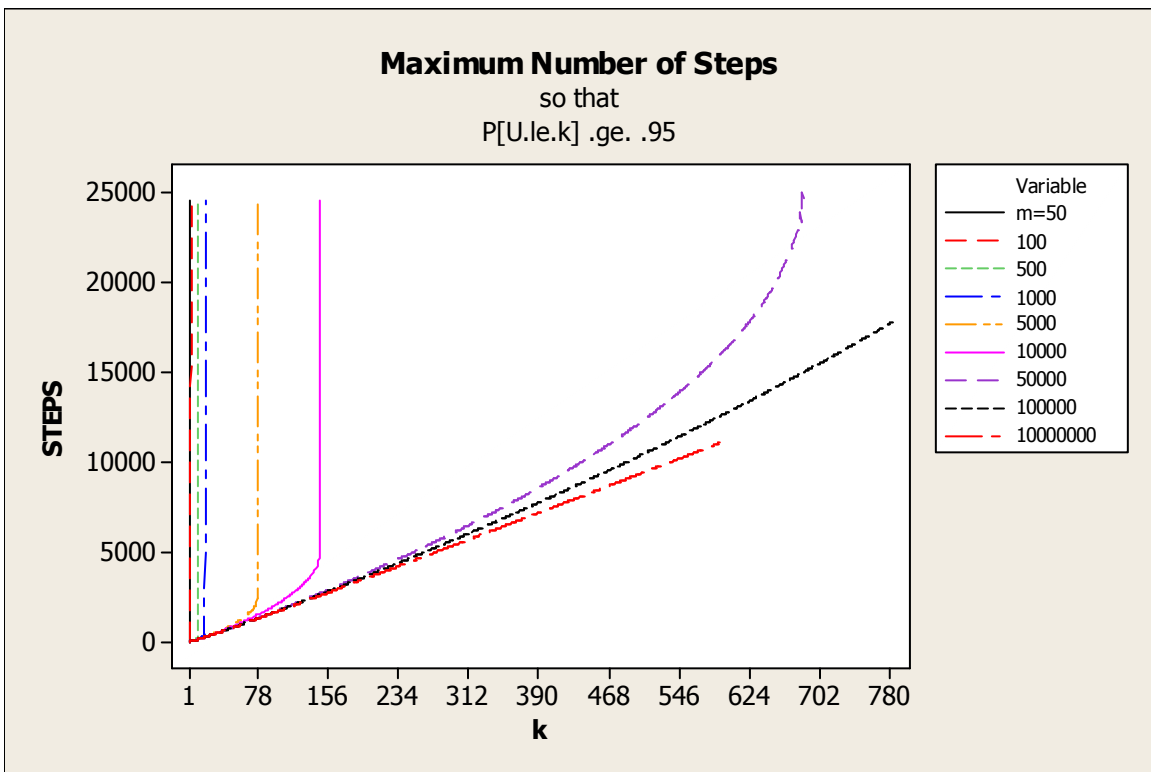


Figure 6.2

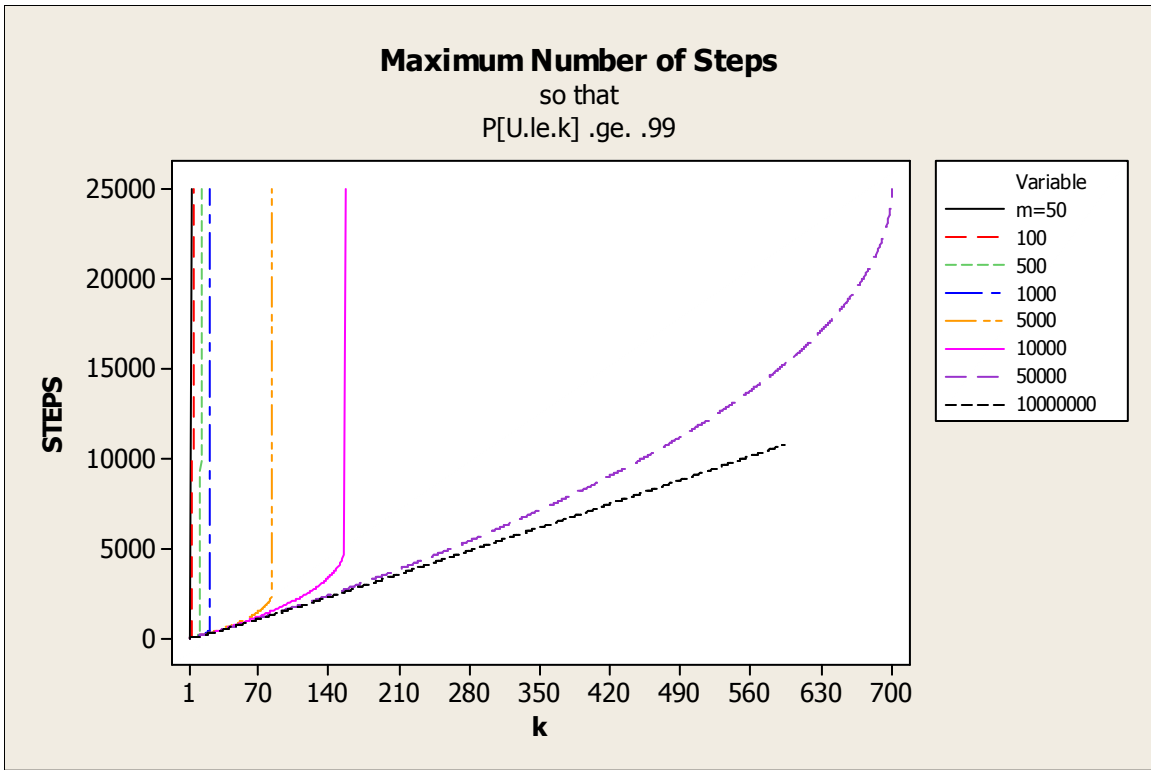


Figure 6.3

It may be observed that for k sufficiently large, there appears to be no restriction on the number of steps s for which $P[U \leq k]$ exceeds a given confidence level. Table 6.1 gives the smallest value of k for which there is a restriction on the number of steps such that $P[U \leq k]$ is greater than or equal to .90, .95 and .99, for various values of m ranging from 50 to 50,000 ($q = .05$).

m	CONFIDENCE					
	.90		.95		.99	
k	s	k	s	k	s	
50	1 (14)	1 (8)	2 (10)			
100	2 (30)	2 (19)	4 (40)			
500	9 (204)	10 (198)	12 (191)			
1000	17 (468)	18 (411)	21 (422)			
5000	74 (2422)	77 (2381)	83 (2389)			
10000	142 (4746)	146 (4662)	155 (4887)			
50000	673 (24551)	683 (24970)	701 (25050)			

Table 6.1

Largest value of k such that the number of steps s in the B-H procedure is limited by the number of steps for $P[U < k] \geq \text{CONF}$ the B-H procedure when $q = .05$. Also, (in parentheses), maximum number of steps for which the inequality holds.

Table 6.2 gives the maximum number of steps such that $P[U \leq k] \geq .90, .95$ and .99 for various values of m ranging from 1 to 10,000,000.

$k \backslash m$	50	100	500	1000	10,000	10,000,000
1	14(8)3	11(7)3	10(7)2	10(7)2	10(7)2	10(7)2
2	50(50)10	30(19)9	22(16)8	22(16)8	22(16)8	22(15)8
3	50(50)50	100(100)19	37(28)16	36(28)16	35(27)16	34(27)16
4		100(100)40	54(42)26	51(40)26	48(39)25	48(39)25
5		100(100)100	72(58)38	67(55)36	63(52)35	63(52)35
6			94(76)51	84(70)48	78(66)46	77(65)46
7			120(97)66	103(86)61	93(80)58	93(79)58
8			153(121)82	122(103)75	109(94)70	108(93)70
9			204(152)101	143(122)90	125(109)83	124(108)82
10			500(198)124	166(142)105	142(126)96	140(123)95
15			500(500)500	321(269)201	227(204)166	222(200)163
16				373(305)225	245(221)180	239(216)177
17				468(349)252	262(237)195	256(232)192
18				1000(411)282	280(254)211	273(248)206
19				1000(1000)317	298(272)226	290(265)221
20				1000(1000)360	317(289)241	307(281)236
30					506(469)406	482(448)390
40					708(662)583	660(621)551
50					922(868)773	841(797)717
60					1150(1087)976	1024(975)886
70					1393(1320)1192	1208(1154)1057
80					1655(1570)1423	1394(1335)1231
90					1939(1841)1671	1599(1335)1231
100					2253(2138)1941	1767(1701)1582
500					10000(10000)10000	9459(9303)9015

Maximum Number of Steps Such That $P[U \leq k] \geq .90 (.95) .99$ ($q = .05$)

Table 6.2

7. SUMMARY AND CONCLUSIONS

The procedure of Somerville and Hemmelmann (2008) is applied to the Benjamini-Hochberg procedure. Tables and graphs are presented which give the maximum number of steps which can be used so that the number of false positives is less than a prescribed number with confidence levels of .90, .95 and .99.

It may be noted that reduced step procedures are simple and easy to apply. The number of steps may be arbitrarily set, or can be determined either by setting an MCV, or setting a maximum p-value for rejection of an hypothesis.

While large values for s increase power, simultaneous increases in the FDR occur. Simulation studies by Somerville (2004) suggested that using larger values for s than the actual number of false hypotheses is counterproductive, since the increase in power, if any, is “negligible”. In addition, increasing s increases the probability that hypotheses with larger p-values will be rejected. Although knowledge of the number of false hypotheses is usually limited, strict limitation of the number of steps (using reduced step procedures) seems prudent.

8. REFERENCES

Benjamini Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J.R.Statist. Soc. B*, 289-300.

Lehmann, E. L. and Romano, Joseph P. (2005). Generalizations of the familywise error rate. *Annals of Statistics*, **33**, No. 3, 1138-1154.

Somerville, Paul N. (2004). FDR step-down and step-up procedures for the correlated case. *Recent Developments in Multiple Comparison Procedures, IMS Lecture Notes - Monograph Series*, **47**, 100-118.

Somerville, Paul N. and Hemmelmann, Claudia (2008) FDR procedures controlling the number and proportion of false positives. *Computational Statistics and Data Analysis* **52**, 1323-1334.

Steingrimsson, Einar (2007). Private communication.

van der Laan, Mark J., Dudoit, Sandrine and Pollard, Katherine S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* **3**, 1-25.

APPENDIX

Theorem:

Given j blocks, j elements, with $k > 0$ elements in the first block, the number of partitions of the remaining $j-k$ elements into the remaining $j-1$ blocks, such that the number of elements in the first i blocks is at least i , is $C(j,k) = k j^{j-k-1}$.

Proof (incomplete):

Let v be the number of element in the second block, and let $b(v,u,j)$ be the number of partitions for given u , v , and j where $u = j-k$. Then, using an odometer like counting process, and Bernouilli polynomials, the following table was obtained for $0 < v \leq u \leq 4$, and $j > 1$.

v	$u=1$	$u=2$	$u=3$	$u=4$
1	1	$2(j-2)$	$3(j-1)(j-3)$	$4(j-1)^2(j-4)$
2		1	$3(j-2)$	$6(j-1)(j-3)$
3			1	$4(j-2)$
4				1

If C_v^u is the number of combinations of u , taken v at a time, then, for the elements in the table,

$$b(v, u, j) = C_v^u (j-1)^{u-v-1} (j+v-u-1).$$

Assume, for the moment, that the formula for $b(v, u, j)$ is valid for all $0 < v \leq u$.

Let w be the number of consecutive blocks, ($0 \leq w \leq k-1$, or $0 \leq w \leq j-u-1$), beginning with block 2, containing no elements. If $w=0$, the number of partitions is

$$\sum_{(v=1,u)} b(v, u, j) = b(\cdot, u, j) \quad (\text{say}) \quad v \leq u.$$

If $w = 1$, the number of additional partitions is given by

$$\sum_{(v=1,u)} b(v, u, j-1) = b(\cdot, u, j-1) \quad v \leq u.$$

In general, w cannot be greater than $k-1$, and we have

$$\begin{aligned} C(j,k) &= \sum_{(w=0,j-u-1)} b(\cdot, u, j-w) \\ &= \sum_{(w=0,j-u-1)} \sum_{(v=1,u)} C_v^u (j-w-1)^{u-1} (j-w+v-u-1). \end{aligned}$$

A Fortran95 program to calculate $C(j,k)$ has been written, assuming the formula for $b(v, u, j)$ holds for all $0 < v \leq u$, and a number of calculations were made. The calculations in all cases (including $u=j-k=1, 2, 3$, and 4) gave

$$C(j,k) = \sum_{(w=0,j-u-1)} b(\cdot, u, j-w) = k j^{j-k-1}, \text{ or } (j-u) j^{u-1}.$$

The problem has also been investigated by Steingrímsson (2007), who observed that proof of the above theorem is equivalent to proof of the identity

$$k \sum_i C_i^{n-k} n^i = \sum_i (n-i) C_i^{n-k+1} n^{i-1}.$$

Using standard MAPLE summation tools, he found the identity to be true.

It is worth mentioning that the formula for A_{f+i} in section 5., was used for $m=5$ and $m=20$, to calculate the value of A_{f+i} for $q=.05$, and for all possible values of s , f and i . The Fortran95 program FDRpwr.F95 was also used to calculate A_{f+i} for the same parameter values. In all cases the calculated values were the same, rounded to 3 decimal places. The program FDRpwr uses Monte Carlo simulation. The program can be used to obtain, in addition to A_{f+i} , three kinds of power, probabilities and cumulative probabilities of rejections of at most k true hypotheses and cumulative probabilities of rejection of at least k hypotheses (arbitrary k).