Technical Report 1004
Dept. of Biostatistics

Some Exact and Approximations for the Distribution of the Realized False Discovery Rate

David Gold[ab], Jeffrey C. Miecznikowski[ab1]

[a]Department of Biostatistics, University at Buffalo, Buffalo, NY 14214-3000, USA
[b]Department of Biostatistics, Roswell Park Cancer Institute, New York 14263

Short title:

Realized False Discovery Rate

Proofs to be sent to:

Jeffrey C. Miecznikowski
Department of Biostatistics
School of Public Health and Health Professions
249 Farber Hall
University at Buffalo
3435 Main Street
Buffalo NY 14214-3000, USA

[1]Corresponding Author. Department of Biostatistics, School of Public Health and Health Professions, 249 Farber Hall, University at Buffalo, 3435 Main Street, Buffalo NY 14214-3000, USA. Tel:+1(716) 829 2754. e-mail:jcm38@buffalo.edu

Some Exact and Approximations for the Distribution of the Realized False Discovery Rate

by David Gold and Jeff Miecznikowski

Here, we derive the distribution of the realized false discovery rate (rFDR), for Benjamini and Hochberg's (1995) procedure, given a general distribution of test statistics.

The following notation is refered to

1. a the desired FDR
2. test statistics, X iid $\pi_0 F_0(X) + \pi_1 F_1(X)$, $X = (X_1, ..., X_m)'$ mixture of the CDF's F_0, the null distribution, and F_1 the alternative, with mixture weights $\pi_0 + \pi_1 = 1$. We are mainly interested in the case of a t-test, where F is a t distribution with v degrees of freedom and F_1 is a (possibly) mixture of non-central t's, with v degrees of freedom and noncentrality parameter η.
3. two-sided p-values $p_i = 2(1 - F_0(|X_i|))$, $i = 1, ..., m$ with distribution $P(p \leq c) = \pi_0 P_0(p \leq c) + \pi_1 P_1(p \leq c)$
4. ordered p-values $p_{(1)}, ..., p_{(m)}$
5. $c_j = aj/m$ for $j = 1, ..., m$

1 Density of the Ordered p-value chosen

Define the sets:

$$
\begin{aligned}
A_j &= \{p_{(j)} \leq aj/m, p_{(j+1)} > a(j+1)/m, p_{(j+2)} > a(j+2)/m, ..., p_{(m)} > a\} \\
B_j &= \{p_{(j)} > aj/m, p_{(j+1)} > a(j+1)/m, ..., p_{(m)} > a\} \\
C_j &= \{p_{(j+1)} > a(j+1)/m, p_{(j+2)} > a(j+2)/m, ..., p_{(m)} > a\}
\end{aligned}
$$

Then $P(A_j) = P(C_j) - P(B_j)$. Note that for sufficiently large m, large π_1, and entropy between P_0 and P_1, the approximation

$$P(A_j) \approx P(\{p_{(j)} \leq aj/m, p_{(j+1)} > a(j+1)/m\})$$

is efficient.

2 Distribution of Order Statistics

The joint distribution of two order statistics is derived in Casella and Berger.

$$P(p_{(i_1)} \le u_1, p_{(i_2)} \le u_2)$$

define

$$\begin{aligned} U_1 &= \sum I(p_i \le u_1) \\ U_2 &= \sum I(u_1 < p_i \le u_2) \end{aligned}$$

$$\begin{aligned} P(p_{(i_1)} \le u_1, p_{(i_2)} \le u2) &= P(i_1 \le U_1 < i_2, i_2 \le U_1 + U_2 \le n) + P(U_1 \ge i_2) \\ &= \sum_{s1=i_1}^{i_2-1} \sum_{s2=i_2-s_1}^{m-s_1} P(U_1 = s_1, U_2 = s_2) + P(U_1 \ge i_2) \end{aligned}$$

for the general case, where $p_1, ..., p_m$ are not necessarily independent or identically distributed,

$$P(U_1 = s_1, U_2 = s_2) = \\ \sum_{q \in \mathcal{Q}} \int_0^{u_1} \cdots \int_0^{u_1} \int_{u_1}^{u_2} \cdots \int_{u_1}^{u_2} \int_{u_2}^{\infty} \cdots \int_{u_2}^{\infty} P_{p_{q_1},...,p_{q_m}}(p_{q_1}, ..., p_{q_m}) dp_{q_1} \cdots dp_{q_m}$$

for the set $\mathcal{Q} = \{q : (q_1, ..., q_m) \text{ are permutations of } (1, ..., m)\}$, and reducing in the iid case to,

$$\begin{aligned} P(U_1 = s_1, U_2 = s_2) &= \frac{m!}{s_1! s_2! (m - s_1 - s_2)!} [P(p \le u_1)]^{s_1} [P(p \le u_2) - P(p \le u_1)]^{s_2} \times \\ &\quad [1 - P(p \le u_2)]^{m - s_1 - s_2} \end{aligned}$$

The joint CDF of k order statistics is derived in Glueck et. al (2008) for the non-identically distributed case, in particular with two sub-populations. In order to calculate the probability of B_j, or that $k = m - j - 1$ of the largest order statistics are greater than constants $c_1, c_2, ..., c_k$, following the logic in Glueck et al. and some of their notation, for $p_1, ..., p_m$ iid F, the joint CDF of the order statistics, $(p_{(j_1)}, ..., p_{(j_k)})$,

$$
\begin{aligned}
P(\cap_{s=1}^{k}\{p_{(j_s)} \leq c_s\}) &= P(\cap_{s=1}^{k}\{\text{ at least } j_s \text{ of } p_i\text{'s } \leq c_s\}) \\
&= P(\cap_{s=1}^{k}\{I_s \geq j_s\}) \\
&= \sum_{\mathbf{i}\in\mathcal{I}} P(\cap_{s=1}^{k}\{I_s = i_s\}) \\
&= \sum_{\mathbf{i}\in\mathcal{I}} P(\cap_{s=1}^{k+1}\{i_s - i_{s-1} \text{ of } p_i\text{'s } \in (c_{s-1}, c_s]\}) \\
&= \sum_{\mathbf{i}\in\mathcal{I}} m! \prod_{s=1}^{k+1} \frac{[P(p \leq c_s) - P(p \leq c_{s-1})]^{(i_s - i_{s-1})}}{(i_s - i_{s-1})!}
\end{aligned}
$$

given $I_s = \sum_{i=1}^{m} I(p_i \leq c_s)$, so that $I_1 \leq \cdots \leq I_k$, leading to the index set

$$\mathcal{I} = \{\mathbf{i} : 0 = i_0 \leq i_1 \leq \cdots \leq i_k \leq i_{k+1} = m, i_s \geq j_s \text{ all } s \in [1, k]\}$$

We, however, are interested in the joint probability

$$
\begin{aligned}
P(\cap_{s=1}^{k}\{p_{(j_s)} > c_s\}) &= P(\cap_{s=1}^{k}\{\text{at most } j_s - 1 \text{ of } p_i\text{'s } \leq c_j\}) \\
&= P(\cap_{s=1}^{k}\{I_s < j_s\}) \\
&= \sum_{\mathbf{i}\in\mathcal{I}} P(\cap_{s=1}^{k}\{I_s = i_s\}) \\
&= \sum_{\mathbf{i}\in\mathcal{I}} P(\cap_{s=1}^{k+1}\{i_s - i_{s-1} \text{ of } p_i\text{'s } \in (c_{s-1}, c_s]\}) \\
&= \sum_{\mathbf{i}\in\mathcal{I}} m! \prod_{s=1}^{k+1} \frac{[P(p \leq c_s) - P(p \leq c_{s-1})]^{(i_s - i_{s-1})}}{(i_s - i_{s-1})!}
\end{aligned}
$$

$$\mathcal{I} = \{\mathbf{i} : 0 = i_0 \leq i_1 \leq \cdots \leq i_k \leq i_{k+1} = m, i_s < j_s \text{ all } s \in [1, k]\}$$

note that for 2-sided p-values, the $P(p \leq c_s) = P(X \leq -X_{c_s}) + P(X \geq X_{c_s})$, where and $P(X \leq -X_{c_s}) = F(-X_{c_s})$, etc., and $X_{c_s} = -F_0^{-1}(.5c_s)$.

The results in Glueck can be extended for a multivariate distribution $P(\cap_{s=1}^{k+1}\{i_s - i_{s-1} \text{ of } p_i\text{'s } \in (c_{s-1}, c_s]\}) =$

$$\sum_{q\in\mathcal{Q}} \int_{c_0}^{c_1} \cdots \int_{c_0}^{c_1} \int_{c_1}^{c_2} \cdots \int_{c_1}^{c_2} \int_{c_k}^{c_{k+1}} \cdots \int_{c_k}^{c_{k+1}} P_{p_{q_1},...,p_{q_m}}(p_{q_1}, ..., p_{q_m}) dp_{q_1} \cdots dp_{q_m}$$

There are i_1 integrals from $(-\infty, c_1), (i_2 - i_1)$ from (c_1, c_2), ..., and $m - i_k$ integrals from (c_k, ∞). In order to compute the respective n-dim integral over 2-sided p-values, for each permutation, there are 2^m possible ways to integrate over the distribution of test statistics, over positive and negative domains, respectively. Suppose for example that $X \sim f_0$ and $X \sim f_1$ are independent, and that within each sub-population, the covariance matrix is block diagonal, with B_0 and B_1 blocks, repsectively. This leads to considerable reductions in computation, i.e. the product of B_0- and B_1-dim integrals, rather than one m-dim integral, assuming the order of integration is inter-changable. Permuations within-block are not necessary to compute, where variables are exchangable.

3 Density of the p-value threshold

$$P(\tau) = P(0) + \sum_{j=1}^{m} P(\tau|j)P(A_j)$$

where $P(0)$ is the probability that no tests are rejected, and

$$\begin{aligned} P(\tau|j) &= P(p_{(j)}|A_j) \\ &= \int_{a(j+1)/m}^{1} \cdots \int_{a}^{1} P(p_{(j)}, p_{(j+1)}, ..., p_{(m)}|A_j) dp_{(j+1)} \cdots dp_{(m)} \end{aligned}$$

or,

$$\begin{aligned} P(\tau|j) &= \frac{\partial}{\partial \tau} P(p_{(j)} \leq \tau | A_j) \\ &= \frac{\partial}{\partial \tau} P(p_{(j)} \leq \tau, p_{(j+1)} > a(j+1)/m, ..., p_{(m)} > a)/P(A_j) \end{aligned}$$

if $\tau \leq aj/m$, and 0 otherwise. For the bivariate approximation, let

$$\tilde{A}_j = \{p_{(j)} \leq aj/m, p_{(j+1)} > a(j+1)/m\}.$$

$$
\begin{aligned}
P(\tau|j) &\approx \frac{\partial}{\partial\tau}P(p_{(j)} \leq \tau|\tilde{A}_j) \\
&= \frac{\partial}{\partial\tau}P(p_{(j)} \leq \tau, p_{(j+1)} > a(j+1)/m)/P(\tilde{A}_j) \\
&= \frac{\partial}{\partial\tau}P(p_{(j)} \leq \tau)/P(\tilde{A}_j) - \frac{\partial}{\partial\tau}P(p_{(j)} \leq \tau, p_{(j+1)} \leq a(j+1)/m)/P(\tilde{A}_j)
\end{aligned}
$$

if $\tau < aj/m$ and 0 otherwise. Further, for the iid case,

$$
\frac{\partial}{\partial\tau}P(p_{(j)} \leq \tau) = j\binom{m}{j}P(p=\tau)[P(p \leq \tau)]^{j-1}[1-P(p \leq \tau)]^{m-j}
$$

and

$$
\frac{\partial}{\partial\tau}P(p_{(j)} \leq \tau, p_{(j+1)} \leq a(j+1)/m) = \sum_{s1=j}^{j+1-1}\sum_{s2=j+1-s_1}^{m-s_1}\frac{\partial}{\partial\tau}P(U_1=s_1, U_2=s_2) + \frac{\partial}{\partial\tau}P(U_1 \geq j+1)
$$

where, the partial of $P(U_1 \geq j+1)$ w.r.t τ is found as above, as

$$
(j+1)\binom{m}{j+1}P(p=\tau)[P(p \leq \tau)]^{(j+1)-1}[1-P(p \leq \tau)]^{m-(j+1)}
$$

and

$$
\begin{aligned}
\frac{\partial}{\partial\tau} & P(U_1=s_1, U_2=s_2) = \\
& \frac{\partial}{\partial\tau}\frac{m!}{s_1!s_2!(m-s_1-s_2)!}[P(p \leq \tau)]^{s_1}[P(p \leq a(j+1)/m) - P(p \leq \tau)]^{s_2} \times \\
& [1-P(p \leq a(j+1)/m)]^{m-s_1-s_2} \\
= & \frac{m!}{s_1!s_2!(m-s_1-s_2)!}[1-P(p \leq a(j+1)/m)]^{m-s_1-s_2} \times \\
& [s_1P(p \leq \tau)^{s_1-1}P(p=\tau)[P(p \leq a(j+1)/m) - P(p \leq \tau)]^{s_2} - \\
& P(p \leq \tau)^{s_1}(s_2)[P(p \leq a(j+1)/m) - P(p \leq \tau)]^{s_2-1}P(p=\tau)]
\end{aligned}
$$

Again, note $P(p \leq c)$ is found above.

4 CDF of rFDR

$$rFDR = \begin{cases} \dfrac{w_0}{w_0 + w_1} & \text{if } w_0 + w_1 > 0 \\ 0 & \text{if } w_0 + w_1 = 0 \end{cases}$$

where w_0 is the count of false, and w_1 true rejections, respectively. The CDF is defined as

$$P(rFDR \leq c|j) \quad = \quad \sum_{w_0, w_1 : \, rFDR \leq c} \sum P(w_0, w_1|m_0, m_1, F, j)P(m_0, m_1|m, F)$$

Stating the obvious,

$$\begin{aligned} m_0 &\sim Binom(m, \pi_0) \\ m_1 &= m - m_0 \end{aligned}$$

We can find the conditional joint distribution

$$P(w_0, w_1|m_0, m_1, F, j) = P(w_0, w_1, j|m_0, m_1, F)/P(j|m_0, m_1, F)$$

as described at the end of Section 3, letting f be partitioned into two sub-populations of size m_0 and m_1, requiring that w_0 and w_1 of the integration limits be $(0, c_1)$ respectively by sub-population. Also, consider the results in Glueck for two populations.

Then we have

$$P(rFDR \leq c) = \sum_j P(rFDR \leq c|j)P(A_j)$$

mote that $P(rFDR \leq c|\tau)$ can be approximated well, for large m, treating w_0, w_1 as independent, e.g. $P(w_0|\tau) \approx \sum_{r_0=0}^{w_0} Binom(r_0, m, \pi_0 P_0(p \leq \tau))$.

$$P(rFDR \leq c) = \int_0^1 P(rFDR \leq c|\tau)P(\tau)d\tau$$

4.1 Independent Case

Under the integral, there is a double sum of independent terms, with each depending on τ. Integrating each term and aggregating yields the result. The sum is composed of the terms

$$
\begin{aligned}
& P(rFDR \leq c|\tau)P(\tau|j) \\
& = \sum_{w_0,w_1:} \sum_{rFDR\leq c} \sum_{r_0=1}^{w_0} \binom{m}{r_0} (\pi_0\tau)^{r_0}(1-\pi_0\tau)^{m-r_0} \cdot \\
& \sum_{r_1=1}^{w_1} \binom{m}{r_1} (\pi_1\tau)^{r_1}(1-\pi_1 F_1(\tau))^{m-r_1} \cdot \\
& \left[\frac{\partial}{\partial\tau} P(p_{(j)} \leq \tau)/P(\tilde{A}_j) - \frac{\partial}{\partial\tau} P(p_{(j)} \leq \tau, p_{(j+1)} \leq a(j+1)/m)/P(\tilde{A}_j)\right]
\end{aligned}
$$

Then for any r_0, r_1, in the above,

$$
= E_1 \cdot \left[E_2 - \left[E_3 + \sum_{s1=j}^{j+1-1} \sum_{s2=j+1-s_1}^{m-s_1} E_4(s_1,s_2)(E_5(s_1,s_2) - E_6(s_1,s_2))\right]\right] /E_6
$$

where

$$
\begin{aligned}
E_1 &= \binom{m}{r_0}(\pi_0\tau)^{r_0}(1-\pi_0\tau)^{m-r_0}\binom{m}{r_1}(\pi_1\tau)^{r_1}(1-\pi_1 P_1(p\leq\tau))^{m-r_1} \\
E_2 &= j\binom{m}{j}P(p=\tau)[P(p\leq\tau)]^{j-1}[1-P(p\leq\tau)]^{n-j} \\
E_3 &= (j+1)\binom{m}{j+1}P(p=\tau)[P(p\leq\tau)]^{(j+1)-1}[1-P(p\leq\tau)]^{n-(j+1)} \\
E_4 &= \frac{m!}{s_1!s_2!(m-s_1-s_2)!}[1-P(p\leq a(j+1)/m)]^{m-s_1-s_2} \\
E_5 &= s_1 P(p\leq\tau)^{s_1-1}P(p=\tau)[P(p\leq a(j+1)/m)-P(p\leq\tau)]^{s_2-s_1} \\
E_6 &= P(p\leq\tau)^{s_1}(s_2-s_1)[P(p\leq a(j+1)/m)-P(p\leq\tau)]^{s_2-s_1-1}P(p\leq\tau) \\
E_7 &= P(\tilde{A}_j)
\end{aligned}
$$

where f, F are the pdf and cdf of the p-values, respectively. the integrals that need to be performed are proportional to

$$
\int_0^{c_j} \tau^{r_0+r_1}(1-\pi_0\tau)^{m-r_0}(1-\pi_1 P_1(p\leq\tau))^{m-r_1}P(p=\tau)[P(p\leq\tau)]^{j-1}[1-P(p\leq\tau)]^{m-j}d\tau
$$

$$\int_0^{c_j} \tau^{r_0+r_1}(1-\pi_0\tau)^{m-r_0}(1-\pi_1 P_1(p\le\tau))^{m-r_1} P(p=\tau)^{s_1-1} \times$$
$$P(p=\tau)[P(p\le a(j+1)/m) - P(p=\tau)]^{s_2-s_1} d\tau$$

$$\int_0^{c_j} \tau^{r_0+r_1}(1-\pi_0\tau)^{m-r_0}(1-\pi_1 P_1(p\le\tau))^{m-r_1} P(p\le\tau)^{s_1}(s_2-s_1) \times$$
$$[P(p\le a(j+1)/m) - P(p\le\tau)]^{s_2-s_1-1} P(p=\tau) d\tau$$

4.2 Correlation Case

Correlation introduces complexities, that are beyond our capacity, with current state of the art computing. It is unpractical and unrealistic to expect that we will generate results for the general correlated case. However, for restricted and special cases, results can be achieved quickly. One such case is the block diagonal correlation matrix, of blocks of size B, identically distributted in a block, and further assuming that variables following either component distribution f_0 or f_1 are independent. For the approximation, relying on the set $\tilde{A}$, rather than the full set A, we need perform calculations for the joint density of two order statistics. There are four combinations of two cases that we must consider, for two variables that are indepedent or dependent, and belonging to components f_0 or f_1, respectively, and weight results accordingly. For the independent variables, we take previous results. For dependent variables, we compute, for each component weighting accordingly,

$$P(U_1 = s_1, U_2 = s_2) =$$
$$B! \int_0^{u_1} \cdots \int_0^{u_1} \int_{u_1}^{u_2} \cdots \int_{u_1}^{u_2} \int_{u_2}^{\infty} \cdots \int_{u_2}^{\infty} P_{p_1,\ldots,p_B}(p_1,...,p_B) dp_{q_1} \cdots dp_{q_m}$$

where $B!$ is the number of ways to permute B variables. If the block sizes vary, then we may compute over each size, and weight accordingly. If allow variables from each component in a block, then we must weight accordingly, with the correct number of permutations, which must be mixed over the correct binomial distribution, given the population rates. All of these considerations are for the sake of computational speed.

5 SIMULATIONS

To Be Determined

6 REFERENCES

1. Benjamini and Hochberg (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B. 57:289300.

2. Glueck, Karimpour-Fard, Mandel,Hunter, Muller (2008) Fast Computation by Block Permanents of Cumulative Distribution Functions of Order Statistics from Several Populations. Commun Stat Theory Methods. 37(18):28152824. doi: 10.1080/03610920802001896.

3. Casella and Berger (2001) Statistical Inference. Duxbury Press.