A Brief Summary of *Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing*[1]

Hongyi Zheng, Charlie Chen

December 10, 2022

1 Multiple Testing Problems

When we are trying to conduct a statistical analysis task involving multiple simultaneous statistical tests, conducting every single test separately without considering the entire multiple testing problem as a whole may result in some consequences. For instance, suppose we performed 100 independent tests each at level $\alpha = 0.05$, and all null hypotheses are true. The expected number of false rejections is 5, and the probability that at least one null hypothesis is incorrectly rejected is about 99.4%. Therefore, we need some multiple-comparison procedures (MCPs) to control this selection effect.

The first idea is to control the familywise error rate (FWER), which is the probability of rejecting at least one true hypothesis H_i. If m independent tests against m true null hypotheses are performed, and each individual test is conducted at level α, then the family-wise error rate (FWER) is given by

$$\bar{\alpha} = 1 - (1 - \alpha)^m \tag{1}$$

Even if the tests are not independent of each other, let $p_1, p_2, \cdots, p_m$ be the p-values of these m tests, by Boole's inequality, we are still able to conclude that

$$\bar{\alpha} = \mathbb{P}\left\{\bigcup_{i=1}^{m} (p_i \leqslant \alpha)\right\} \leqslant \sum_{i=1}^{m} \{\mathbb{P}(p_i \leqslant \alpha)\} = m\alpha \tag{2}$$

as the p-values are uniformly distributed under the null hypothesis. Thus, if we control the level of each individual test at α/m, then FWER is guaranteed to be at most α. This is also known as the Bonferroni correction.

However, such FWER controlling MCPs has some significant drawbacks, as the conservative test level increases the probability of committing type-II errors and thus reduces the power of the test. Also, often it is not desirable to control the FWER. For instance, considering a treatment and multiple tests for various aspects of the effect. Even if some of the null hypotheses are falsely rejected, it does not necessarily mean that the conclusion of this treatment being effective is erroneous. Therefore, an alternative measure, the False Discovery Rate (FDR), is proposed to address these issues.

2 Introduction to False Discovery Rate

First of all, consider the following table of the testing results:

	Declared non-significant	Declared Significant	Total
True null hypothesis	**U**	**V**	m_0
Non-true null hypothesis	**T**	**S**	$m - m_0$
	$m - \mathbf{R}$	**R**	m

Here **R** is an observable random variable, while **U**, **V**, **S** and **T** are four unobservable random variables. In the following summary, we use the lowercase letter for their realized values. It is not difficult to observe that the per-comparison error rate (PCER), which is the expected proportion of false-positive tests, is $\mathbb{E}(\mathbf{V}/m)$, and the FWER is $\mathbb{P}(\mathbf{V} \geq 1)$.

Now define $\mathbf{Q} = \mathbf{V}/(\mathbf{V} + \mathbf{S})$, the proportion of the rejected null hypothesis which is incorrectly rejected, and define $\mathbf{Q} = 0$ when $\mathbf{V} + \mathbf{S} = 0$. Notice that **Q** is an unobserved random variable as both **V** and **S** are unobservable. We define FDR, Q_e. to be the expectation of **Q**.

$$Q_{\mathrm{e}} = \mathbb{E}(\mathbf{Q}) = \mathbb{E}\{\mathbf{V}/(\mathbf{V} + \mathbf{S})\} = \mathbb{E}(\mathbf{V}/\mathbf{R}) \tag{3}$$

Two important observations:

- If $m_0 = m$ (e.g. all null hypothesis is true), then $\mathbf{Q} = 1$ if $\mathbf{V} > 0$ and $\mathbf{Q} = 0$ if $\mathbf{V} = 0$. Thus in this case

  $$Q_e = \mathbb{E}(\mathbf{Q}) = \mathbb{E}(\mathbb{1}_{\mathbf{V}>0}) = \mathbb{P}(\mathbf{V} > 0) \tag{4}$$

  which means FDR is equivalent to FWER when all null hypotheses are true.

- If $m_0 < m$, then we again have $\mathbf{Q} = 0$ if $\mathbf{V} = 0$, but $\mathbf{Q} \leqslant 1$ if $\mathbf{V} > 0$ because **S** could now be positive integer. Therefore now $\mathbf{Q} \leqslant \mathbb{1}_{\mathbf{V}>0}$, so we have

  $$Q_e = \mathbb{E}(\mathbf{Q}) \leqslant \mathbb{E}(\mathbb{1}_{\mathbf{V}>0}) = \mathbb{P}(\mathbf{V} > 0) \tag{5}$$

  which means FDR is less than or equal to FWER when not all null hypotheses are true. Also we see that the larger **S** tends to be, the larger the difference between **Q** and $\mathbb{1}_{\mathbf{V}>0}$ is. Therefore, when more of the hypotheses are not true, there is a larger potential for an increase in testing power.

3 False Discovery Rate Controlling Procedure

3.1 The FDR Controlling Procedure

Consider testing a set of hypothesis $H_1, H_2, \ldots, H_m$ and the the corresponding p-values $P_1, P_2, \ldots, P_m$. Order these p-values and denote the ordered set as $P_{(1)} \leqslant P_{(2)} \leqslant \ldots \leqslant P_{(m)}$ and rearrange the corresponding hypothesis and denote as $H_{(1)} \leqslant H_{(2)} \leqslant \ldots \leqslant$

$H_{(m)}$. The False Discovery Rate controlling procedure is defined as a Bonferroni-type multiple-testing procedure:

Algorithm 1 FDR Controlling Procedure

Let k be the largest i for which $P_{(i)} \leqslant \frac{i}{m} q^*$,
then we reject all $H_{(i)}$ for $i = 1, 2, \ldots, k$

To show that this procedure controls the False Discovery Rate, we will first show an important lemma

Lemma 3.1. *For any $0 \leqslant m_0 \leqslant m$ independent p-values corresponding to true null hypotheses, there are $m_1 = m - m_0$ p-values corresponding to the false null hypotheses. By the procedure of Algorithm 1, it is true that the expected value of $\mathbf{Q}$ conditioned on the p-values of all the false null hypotheses is less than the targeted FDR reweighted by $\frac{m_0}{m}$. To be specific,*

$$\mathbb{E}\left(\mathbf{Q} \mid P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) \leqslant \frac{m_0}{m} q^* \tag{6}$$

Proof. We use induction on m to prove this theorem.

Base Case

Consider $m = 1$, then this is the single testing scenario. If $m_0 = 0$, then by definition

$$\mathbb{E}(\mathbf{Q} \mid P_1 = p_1) = 0 \leqslant \frac{m_0}{m} q* \tag{7}$$

and if $m_0 = m = 1$, this single true hypothesis is rejected if its corresponding p-value $P \leqslant q^*$, so by 5 and the definition of p-value,

$$\mathbb{E}(\mathbf{Q}) = \mathbb{E}(\mathbb{1}_{\mathbf{V}>0}) = \mathbb{P}(P \leqslant q*) \leqslant q^* = \frac{m_0}{m} q* \tag{8}$$

Inductive Case

Assuming the lemma is true for any $m' \leqslant m$, want to show it is also true for $m + 1$.

Case 1: $m_0 = 0$
That is, all null hypotheses are false. In this case,

$$\mathbb{E}(\mathbf{Q} | P_1 = p_1, \cdots, P_m = p_m) = 0 \leqslant \frac{m_0}{m+1} q^* \tag{9}$$

Case 2: $m_0 > 0$
Let $P_i', i = 1, 2, \ldots, m_0$ as the p-value corresponding to the true null hypotheses and denote the largest of these by P'_{m_0}. For the m_1 p-values for the false null hypotheses, sort and rearrange them as $p_1 \leqslant p_2 \leqslant \ldots \leqslant p_{m_1}$. Denote j_0 to be the largest $0 \leqslant j \leqslant m_1$ satisfying

$$p_j \leqslant \frac{m_0 + j}{m+1} q^* \tag{10}$$

and define

$$p'' \equiv \frac{m_0 + j_0}{m+1} q^* \tag{11}$$

Now, $P'_{(m_0)}$ is the largest p-value for all true null hypotheses, and we denote its probability distribution function as $f'_{P'_{(m_0)}}(p)$. Given that under the null hypothesis, $P'_i \sim \text{Unif}(0,1)$ and P'_i's are independent of each other, the probability density function of $P'_{(m_0)}$ is just

$$f'_{P'_{(m_0)}}(p) = m_0 p^{(m_0-1)} \tag{12}$$

Thus, conditioning on $P'_{(m_0)} = p$, we have

$$\begin{aligned} \mathbb{E}\left(\mathbf{Q} \mid P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) = &\int_0^{p''} \mathbb{E}\left(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, \right. \\ &\left. P_m = p_{m_1}\right) f'_{P'_{(m_0)}}(p)\mathrm{d}p \\ &+ \int_{p''}^1 \mathbb{E}\left(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, \right. \\ &\left. P_m = p_{m_1}\right) f'_{P'_{(m_0)}}(p)\mathrm{d}p \end{aligned} \tag{13}$$

In the first integral, since $p \leqslant p''$, all m_0 true null hypotheses along with j_0 false null hypotheses are rejected, so we know $\mathbf{Q} = m_0/(m_0 + j_0)$ always holds, independent of p. Combining with the definition in Equation 11, the first integral is

$$\frac{m_0}{m_0 + j_0} \left(p''\right)^{m_0} \leqslant \frac{m_0}{m_0 + j_0} \frac{m_0 + j_0}{m+1} q^* \left(p''\right)^{m_0-1} = \frac{m_0}{m+1} q^* \left(p''\right)^{m_0-1} \tag{14}$$

For the second integral, we can dissemble it into segments

$$\begin{aligned} &\int_{p''}^1 \mathbb{E}\left(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) f'_{P'_{(m_0)}}(p)\mathrm{d}p \\ = &\sum_{j_0 \leqslant j \leqslant m_1 - 1} \int_{\max\{p_j, p''\}}^{p_{j+1}} \mathbb{E}\left(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) f'_{P'_{(m_0)}}(p)\mathrm{d}p \end{aligned} \tag{15}$$

By definition of j_0 and p'', in each segment, the $m_1 - j + 1$ hypotheses associated with p-values $p, p_{j+1}, \ldots, p_{m_1}$ cannot be rejected, as $p > p''$ and $p_{j+1}, \ldots, p_{m_1} \geq p_{j_0+1} > p > p''$. Therefore, when all p-values are sorted, $H_{(i)}$ can be rejected only if there is some k satisfying $p_{(k)} \leqslant \{k/(m+1)\}q^*$ and $i \leqslant k \leqslant m_0 + j - 1$ (since we know that $m_1 - j + 1 = m - (m_0 + j - 1)$ hypotheses could not be rejected). Equivalently, we have

$$\frac{p_{(k)}}{p} \leqslant \frac{k}{(m+1)p} q^* = \frac{k}{m_0 + j - 1} \frac{m_0 + j - 1}{(m+1)p} q^* \tag{16}$$

Conditioning on $P'_{(m_0)} = p$, P'_i/p for $i = 1, 2, \ldots, m_0 - 1$ are distributed as $m_0 - 1$ independent $\text{Unif}(0,1)$ random variables, and p_i/p for $i = 1, 2, \ldots, j$ are the smallest

j p-values corresponding to j false null hypotheses, whose values are between 0 and 1 (recall that $p_j \leqslant p$). So, let $\tilde{q^*} = \{m_0 + j - 1/(m+1)p\}q^*$, testing 16 is equivalent to use Algorithm 1 to test $m_0 + j - 1 = m' \leqslant m$ hypotheses at the level $\tilde{q^*}$ with $m_0 - 1$ true null hypotheses. Applying the induction assumption, we have

$$\begin{aligned}\mathbb{E}\left(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) &\leqslant \frac{m_0 - 1}{(m+1)p} q^* \\ &= \frac{m_0 - 1}{m_0 + j - 1} \frac{m_0 + j - 1}{(m+1)p} q^* \\ &= \frac{m_0 - 1}{(m+1)p} q^*\end{aligned} \tag{17}$$

Note that this equation depends on p, but not the segment $p_j \leqslant p \leqslant p_{j+1}$. So the second integral can be bounded as

$$\begin{aligned}&\int_{p''}^{1} \mathbb{E}\left(\mathbf{Q} \mid P'_{(m_0)} = p, P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) f'_{P'_{(m_0)}}(p)\mathrm{d}p \\ \leqslant &\int_{p''}^{1} \frac{m_0 - 1}{(m+1)p} q^* m_0 p^{(m_0-1)} \mathrm{d}p \\ = &\frac{m_0}{m+1} q^* \int_{p''}^{1} (m_0 - 1)\, p^{(m_0-2)} \mathrm{d}p \\ = &\frac{m_0}{m+1} q^* \left\{1 - p''^{(m_0-1)}\right\}\end{aligned} \tag{18}$$

Combining Equation 14 and 18 completes the proof. □

With this lemma being proved, we now state the theorem that guarantees the False Discovery Rate of Algorithm 1.

Theorem 3.2. *For independent test statistics and for any configuration of false null hypotheses, the FDR through Algorithm 1 is bounded by q^*.*

Proof. By Lemma 3.1,

$$\mathbb{E}\left(\mathbf{Q} \mid P_{m_0+1} = p_1, \ldots, P_m = p_{m_1}\right) \leqslant \frac{m_0}{m} q^* \tag{19}$$

Integrate over the joint probability density function of $P_{m_0+1}, \ldots P_m$, we obtain

$$\mathbb{E}\left(\mathbf{Q}\right) \leqslant \frac{m_0}{m} q^* \leqslant q^* \tag{20}$$

Thus FDR is controlled □

3.2 Alternative Look for the FDR Controlling Procedure

To better understand the procedure in Algorithm 1, the author states an alternative look for the controlling procedure. They state that the procedure is equivalent to the following constrained maximization problem.

Theorem 3.3. *The FDR controlling procedure in Algorithm 1 is equivalent to the maximization problem of $r(\alpha)$, which is the number of rejections at level α subject to the constraint $\alpha m/r(\alpha) \leqslant q^*$*

Proof. First note that $p_{(i)} \leqslant \alpha < p_{(i+1)}$ implies $r(\alpha) = i$ and with the same $r(\alpha)$, smaller α decreases the value on the constraint, so it does no harm, if not being beneficial, to always floor α to the closest p-value. Therefore, we only need to investigate α when they equal some p-values. Under the the procedure in Algorithm 1, $\alpha = p_{(k)}$ (meaning rejecting hypotheses corresponding to the k smallest p-values) satisfies the constraint because

$$\frac{\alpha}{r(\alpha)} = \frac{p_{(k)}}{k} \leqslant \frac{q^*}{m} \tag{21}$$

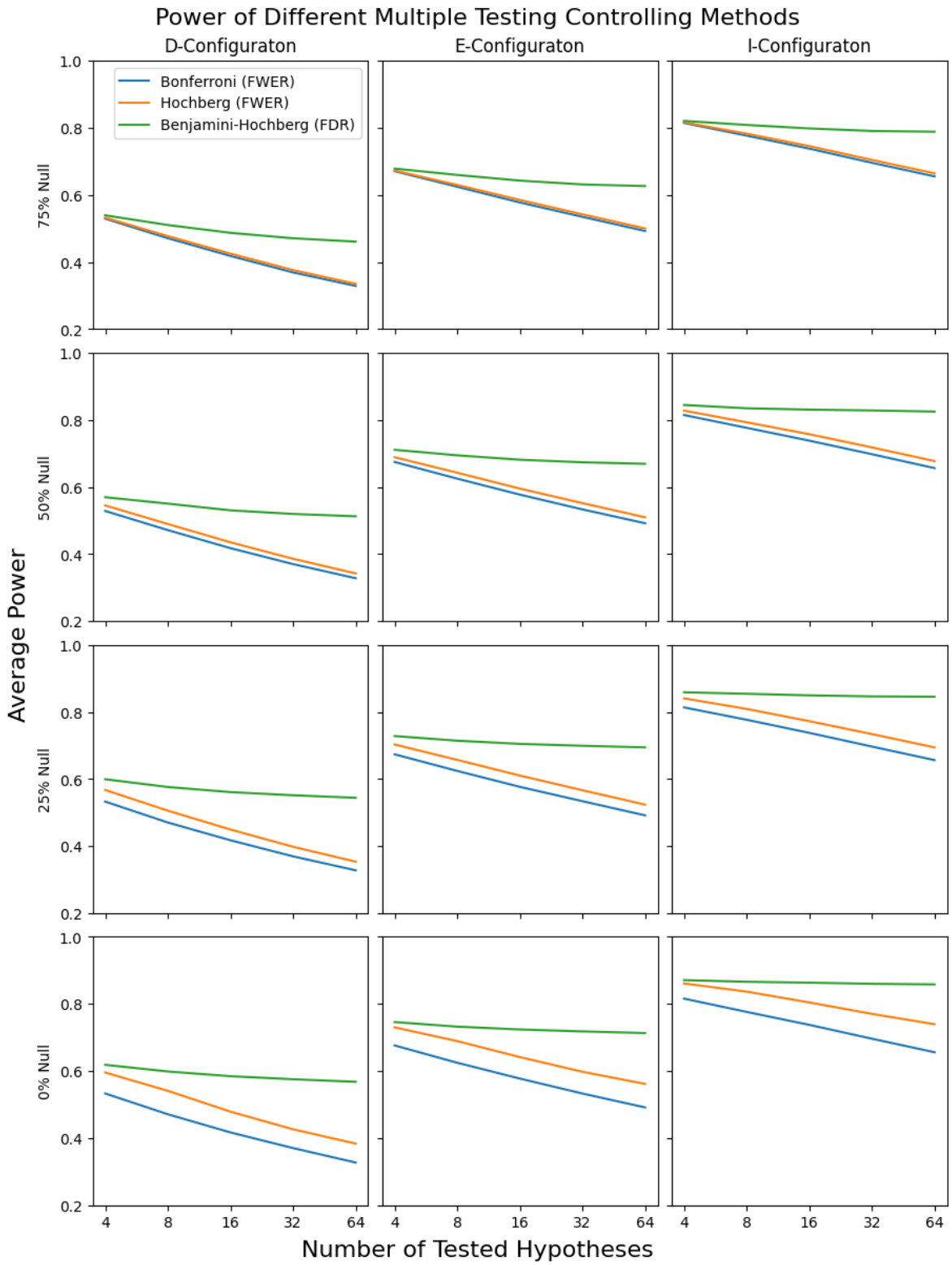
So by considering the largest α first in Algorithm 1 (larger α implies larger $r(\alpha)$), the procedure chooses the α corresponding to the maximum $r(\alpha)$ that satisfies the constraint. □

Therefore, the procedure is balancing between rejecting more hypotheses and controlling FDR. When each individual test has level α, the expected number of wrong rejections is bounded, $\mathbb{E}(\mathbf{V}) \leqslant \alpha m$. After observing the outcome (after we know $r(\alpha)$, the number of rejections made), an upper bound estimate for FDR is $\alpha m/r(\alpha)$. So the Theorem 3.3 suggests that the procedure is maximizing $r(\alpha)$ while controlling FDR under a certain threshold q^*. With this alternative look, we can now appreciate the beauty and delicacy of this controlling procedure!

4 Recreating the Experiment

We recreated the experiment in section 4 of Benjamini and Hochberg's paper, and the results are displayed below. The result slightly differs from the one Benjamini and Hochberg produced, partly due to the ambiguity of the description of the setting in the original paper. It is unclear how to divide non-zero expectations into four groups with linearly increasing / linearly decreasing / equal sizes, especially when m is small and the proportion of the null hypotheses is large as in those cases the total number of non-true hypotheses could be less than four in each configuration.

To address this, we generate observations from non-true hypotheses for all 20000 simulation trials together, and then randomly assign them to each trial. For example, if $m = 8$ and the number of truly null hypotheses is $3m/4 = 6$, then for each trial there would only be 2 nontrue hypotheses and there will be no way to divide these two hypotheses into four groups. Instead, across 20000 trials, we need $20000 \times 2 = 40000$ non-true hypotheses altogether, so we could divide these 40000 hypotheses into four groups with linearly increasing / linearly decreasing / equal sizes and place them at $L/4$, $L/2$, $3L/4$, and L. Then we drew observations from these hypotheses and assign two of them to each trial randomly. This approach allows us to deal with small m and reproduce the experiment.



Despite the slight difference between our experiment and the one done by Benjamini and Hochberg, the following key observations made by Benjamini and Hochberg still hold:

- The power of all the methods decreases when the number of hypotheses tested

increases.

- The power is smallest for the D-configuration, where the non-null hypotheses are closer to the null, and is largest for the I-configuration.
- The power of the FDR controlling method is uniformly larger than that of the other methods.
- The advantage increases with the number of non-null hypotheses.
- The advantage increases in m. Therefore, the loss of power as m increases is relatively small for the FDR controlling method in the E- and I-configurations.
- The gain in power due to the control of the FDR rather than the FWER is much larger than the gain of the FWER controlling method over the Bonferroni method [2].

5 Conclusion

In summary, this FDR controlling procedure proposed by Benjamini and Hochberg provided a powerful testing approach that increases the testing power by a large margin. Compared with the classic conservative approaches which control FWER in a strong sense, this alternative approach controls the FWER in a weak sense through controlling FDR, which could be quite useful in cases where controlling FWER is not of great importance, or where the cost of committing type-II errors are relatively high. In general, this innovative approach from Benjamini and Hochberg significantly reduces the cost paid for the control of multiplicity.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[2] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.