Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 320425, 19 pages
doi:10.1155/2012/320425

Research Article

Control of the False Discovery Proportion for Independently Tested Null Hypotheses

Yongchao Ge[1] and Xiaochun Li[2]

[1] *Department of Neurology, Mount Sinai School of Medicine, One Gustave L. Levy Place, P.O. Box 1137, New York, NY 10029, USA*
[2] *Division of Biostatistics, School of Medicine, New York University, 650 First Avenue, 5th Floor, New York, NY 10016, USA*

Correspondence should be addressed to Yongchao Ge, yongchao.ge@mssm.edu

Received 14 December 2011; Accepted 8 February 2012

Academic Editor: Yongzhao Shao

Consider the multiple testing problem of testing m null hypotheses $H_1, \ldots, H_m$, among which m_0 hypotheses are truly null. Given the P-values for each hypothesis, the question of interest is how to combine the P-values to find out which hypotheses are false nulls and possibly to make a statistical inference on m_0. Benjamini and Hochberg proposed a classical procedure that can control the false discovery rate (FDR). The FDR control is a little bit unsatisfactory in that it only concerns the expectation of the false discovery proportion (FDP). The control of the actual random variable FDP has recently drawn much attention. For any level $1-\alpha$, this paper proposes a procedure to construct an upper prediction bound (UPB) for the FDP for a fixed rejection region. When $1 - \alpha = 50\%$, our procedure is very close to the classical Benjamini and Hochberg procedure. Simultaneous UPBs for all rejection regions' FDPs and the upper confidence bound for the unknown m_0 are presented consequently. This new proposed procedure works for finite samples and hence avoids the slow convergence problem of the asymptotic theory.

1. Introduction

In this paper, we consider the problem of testing m null hypotheses $H_1, \ldots, H_m$, among which m_0 hypotheses are truly null. We shall assume that P-values are available for individual hypotheses. In a seminal paper, Benjamini and Hochberg [1] proposed the false discovery rate (FDR) as an alternative to the classically defined family-wise error rate (FWER). The proposed FDR achieves a good balance between the P-value itself and the FWER correction [2]; the former may give too many false positives, and the latter may give too many false negatives. However, the control of the FDR is a little bit unsatisfactory in that it only concerns the expectation of the false discovery proportion (FDP). In practice, researchers may be

interested in more detailed statistical inference on the actual random variable FDP, not just its expectation. The goal of this paper is to provide a simple procedure to control the FDP.

Let us first introduce some notation. Given m hypotheses $H_1, \ldots, H_m$, let the complete index set be $M = \{1, \ldots, m\}$, M_0 the unknown subset of M for which the null hypotheses are true, and $M_1 = M \setminus M_0$ the subset for which null hypotheses are false. Denote that $m_0 = |M_0|, m_1 = |M_1|$, where $|\cdot|$ denotes the cardinality of a set. The P-values for testing the m hypotheses are $P_1, \ldots, P_m$. A *fixed rejection region* for the P-values can conveniently be taken as $[0, t]$ $(0 < t < 1)$. The value of t could be 0.05, for example. Define the number R_t of all rejected hypotheses and the number V_t of falsely rejected hypotheses, respectively,

$$R_t = \sum_{i=1}^{m} I(P_i \le t), \qquad V_t = \sum_{i \in M_0} I(P_i \le t). \tag{1.1}$$

Following the notation of Korn et al. [3], and Genovese and Wasserman [4], Lehmann and Romano [5], the *false discovery proportion* is defined to be the proportion of falsely rejected null hypotheses among the rejected ones,

$$Q_t = \frac{V_t}{R_t}, \tag{1.2}$$

where the ratio is defined to be zero when the denominator is zero. For a given fixed rejection region $[0, t]$, R_t, V_t, and Q_t are random variables. R, V, and Q will be shorthand for R_t, V_t, and Q_t respectively, when the rejection region $[0, t]$ is clear from the context. The *false discovery rate* of Benjamini and Hochberg [1] is

$$\mathrm{FDR} = E(Q). \tag{1.3}$$

A good understanding of Q will lead investigators to pick an appropriate rejection region $[0, t]$ of P-values. As Q is an unobservable random variable depending on the observed P-values and the rejection region $[0, t]$, the quantity FDR just describes the expectation of this random variable Q. One way to have a more detailed statistical inference on the random variable Q is to derive its distribution, which is very difficult unless a strong assumption can be imposed on the P-values from the false null hypotheses. A conservative approach is to compute an upper prediction bound for Q so that we can safeguard against excessive type I errors. In Section 2, for a fixed rejection region $[0, t]$, we can compute an *upper prediction bound* (UPB) $\overline{Q}_{1-\alpha}(t)$ for Q_t such that

$$\mathrm{pr}\left(Q_t \le \overline{Q}_{1-\alpha}(t)\right) \ge 1 - \alpha. \tag{1.4}$$

If $\overline{Q}_{1-\alpha}(t)$ had been a nonrandom variable, then it should be always no less than the $1 - \alpha$ quantile of the random variable Q_t. When $1 - \alpha = 50\%$, our procedure is very close to the classical BH procedure of Benjamini and Hochberg [1]. In other words, the BH procedure gives us an approximate 50% upper prediction bound (UPB) for Q. With different degrees of being conservative, one should take $1 - \alpha$ at 0.9, 0.95, and 0.99 to ensure high coverage of the false discovery proportion. We also describe how to compute an upper confidence bound

(UCB) for m_0, the number of true null hypotheses. The UCB for m_0 can be used to improve the estimate $\overline{Q}_{1-\alpha}(t)$. In practice, the rejection region $[0,t]$ needs to be adapted to the actual dataset. In Section 3, we give a procedure to construct an upper prediction band for Q_t for all $t \in (0,1)$, and this upper prediction band can be used to pick a data-defined rejection region $[0,\tau]$ of P-values such that the false discovery proportion Q can be controlled at target level γ with prediction accuracy $1-\alpha$, that is,

$$\mathrm{pr}\left(Q_\tau \le \gamma\right) \ge 1-\alpha. \tag{1.5}$$

Thus with probability at least $1-\alpha$, the value of Q is γ or less. For the independent true null P-values, Genovese and Wasserman [4], Meinshausen and Rice [6] also worked on the control of the FDP in the sense of the above equation. However, their results are based on asymptotic theory, while our focus is on the finite-sample results and avoids the slow convergence problem of the asymptotic theory. Other works such as Lehmann and Romano [5], Romano and Shaikh [7, 8], and van der Laan et al. [9] proposed procedures that allow dependence in the P-values but have potentially lost statistical power as the dependence information is not exploited. Section 4 presents a focused statistical inference by restricting the rejection regions onto $\{[0,t] : t \in [t_0, t_0']\}$, which unifies the results of Sections 2 and 3. Section 5 generalizes the results from independent data to less-independent situations, and Section 6 gives our discussion.

2. Finding a $1-\alpha$ UPB for the False Discovery Proportion for a Fixed Rejection Region

For the sake of simplicity, we will first assume that the P-values from the true null hypotheses are following mutually independently uniform distribution $U[0,1]$. We have no further assumptions on the P-values from false null hypotheses. This assumption is the same as in Benjamini and Hochberg [1]. In Section 5 we will generalize the result to less independent situations. For a fixed rejection region $[0,t]$ of the P-values, we would like to find the $1-\alpha$ upper prediction bound (UPB) for the false discovery proportion Q_t. As we mentioned in Section 1, the distribution of Q_t is unknown. However, for any given experimental data, the total number of rejections, R_t, can be easily obtained by (1.1). Under the assumption that true null P-values are independently distributed as $U[0,1]$, V_t has a binomial distribution $\mathrm{Bin}(m_0,t)$. Let U_i, $i = 1,\ldots,m$ be random variables mutually independently distributed as $U[0,1]$, and $N_{m_0,t} = \sum_{i=1}^{m_0} I(U_i \le t)$ distributed as $\mathrm{Bin}(m_0,t)$, hence,

$$V_t \overset{d}{=} N_{m_0,t}. \tag{2.1}$$

The $1-\alpha$ quantile for V_t is the $1-\alpha$ quantile $C_{1-\alpha}(m_0,t)$ of the distribution $\mathrm{Bin}(m_0,t)$. Here $C_{1-\alpha}(m_0,t)$ is defined as

$$C_{1-\alpha}(m_0,t) = \min\left\{k : \mathrm{pr}(N_{m_0,t} \le k) \ge 1-\alpha\right\}. \tag{2.2}$$

As R_t can be computed from the observed data, a $1-\alpha$ UPB for Q_t can be estimated by

$$\overline{Q}_{1-\alpha}(m_0,t)=\frac{C_{1-\alpha}(m_0,t)}{R_t}. \tag{2.3}$$

Lemma 2.1. *For any given* $0\le m_1\le m_2\le m$,

(a) $C_{1-\alpha}(m_1,t)\le C_{1-\alpha}(m_2,t)$,

(b) $m_1-C_{1-\alpha}(m_1,t)\le m_2-C_{1-\alpha}(m_2,t)$, *and*

(c) *let* $g(k)=C_{1-\alpha}(k,t)$ *and* $h(k)=k-C_{1-\alpha}(k,t)$. *The values that* $g(k+1)-g(k)$ *and* $h(k+1)-h(k)$ *take can only be zero or one.*

Proof. By noting that $N_{m_1,t}\le N_{m_2,t}$ when $m_1\le m_2$, we have $\mathrm{pr}(N_{m_1,t}\le k)\ge \mathrm{pr}(N_{m_2,t}\le k)$. Applying the definition of $C_{1-\alpha}$, we obtain the result for part (a).

Note that

$$\begin{aligned}C_{1-\alpha}(m_1,t)&=\min\{k:\mathrm{pr}(N_{m_1,t}\le k)\ge 1-\alpha\}\\&=\min\{k:\mathrm{pr}(m_1-N_{m_1,t}\ge m_1-k)\ge 1-\alpha\}\\&=m_1-\max\{k':\mathrm{pr}(m_1-N_{m_1,t}\ge k')\ge 1-\alpha\}\quad(\text{use } k' \text{ to replace } m_1-k).\end{aligned} \tag{2.4}$$

Therefore,

$$m_1-C_{1-\alpha}(m_1,t)=\max\{k':\mathrm{pr}(m_1-N_{m_1,t}\ge k')\ge 1-\alpha\}. \tag{2.5}$$

Similarly, we can obtain that

$$m_2-C_{1-\alpha}(m_2,t)=\max\{k':\mathrm{pr}(m_2-N_{m_2,t}\ge k')\ge 1-\alpha\}. \tag{2.6}$$

By noting that

$$m_1-N_{m_1,t}=\sum_{i=1}^{m_1}I(U_i>t)\le\sum_{i=1}^{m_2}I(U_i>t)=m_2-N_{m_2,t}, \tag{2.7}$$

we have

$$\mathrm{pr}(m_1-N_{m_1,t}\ge k')\le\mathrm{pr}(m_2-N_{m_2,t}\ge k'). \tag{2.8}$$

Combining this with (2.5) and (2.6) leads to the result for part (b).

Parts (a) and (b), respectively, say that $g(k)$ and $h(k)$ are both increasing functions of k. Simple algebra can establish that $\{g(k+1)-g(k)\}+\{(h(k+1)-h(k)\}=1$. Both $\{g(k+1)-g(k)\}$ and $\{h(k+1)-h(k)\}$ are nonnegative due to the increasing property of functions $g(k)$ and $h(k)$, and hence $0\le g(k+1)-g(k)\le 1$ and $0\le h(k+1)-h(k)\le 1$. The only values that $\{g(k+1)-g(k)\}$ and $\{h(k+1)-h(k)\}$ take can only be zero and one as functions $g(k)$ and $h(k)$ only take integer values. Thus, we complete the proof for part (c). □

Lemma 2.2. *For any given t, $0 < t < 1$, $\overline{Q}_{1-\alpha}(m_0, t)$ of (2.3) is a $1-\alpha$ UPB for the false discovery proportion Q_t, that is,*

$$\mathrm{pr}\Big(Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\Big) \ge 1-\alpha. \tag{2.9}$$

The proof is straightforward by using the fact that $V_t \overset{d}{=} N_{m_0,t}$. We have

$$\begin{aligned}
\mathrm{pr}\Big(Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\Big) &= \mathrm{pr}\left(\frac{V_t}{R_t} \le \frac{C_{1-\alpha}(m_0, t)}{R_t},\ R_t > 0\right) + \mathrm{pr}(R_t = 0) \\
&= \mathrm{pr}(V_t \le C_{1-\alpha}(m_0, t),\ R_t > 0) + \mathrm{pr}(R_t = 0) \\
&= \mathrm{pr}(V_t \le C_{1-\alpha}(m_0, t)) \\
&\ge 1-\alpha.
\end{aligned} \tag{2.10}$$

In the third line, we have used the fact that $\{V_t \le C_{1-\alpha}(m_0, t)$ and $R_t = 0\}$ is the same as the set $\{R_t = 0\}$, which is obtained by noting that V_t must be zero when $R_t = 0$. Following this proof, we can easily see that

$$\left\{Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\right\} = \{V_t \le C_{1-\alpha}(m_0, t)\}. \tag{2.11}$$

The basic construction of $\overline{Q}_{1-\alpha}(m_0, t)$ in (2.3) is the idea central to formulating prediction inference for Q_t. In practice, m_0 is an unknown parameter. The most conservative approach is to replace m_0 with m, in which case we obtain a *conservative* $1-\alpha$ UPB for Q_t. The independence assumption among true null P-values can be used to give a confidence inference for m_0; thus, we can find a better estimate of the UPB for Q_t. For any given $0 < \lambda < 1$, a $1-\alpha$ UCB for m_0 is given by

$$\overline{m}_{0,1-\alpha}(\lambda) = \begin{cases} \max\limits_{k=0,\ldots,m-1}\{k : h(k) = m - R_\lambda\} & \text{if } h(m) < m - R_\lambda, \\ m & \text{otherwise,} \end{cases} \tag{2.12}$$

where $h(k) = k - C_{1-\alpha}(k, \lambda)$ as defined in Lemma 2.1(c). Since $h(0) = 0$ and $h(k+1) - h(k)$ takes value of only zero and one, there exists at least one k, $k \in \{0, \ldots, m-1\}$ such that $h(k) = m - R_\lambda$ when $h(m) < m - R_\lambda$. Therefore, $\overline{m}_{0,1-\alpha}(\lambda)$ in (2.12) is well defined. The parameter λ in (2.12) is used to construct a UCB for m_0; more discussion about it can be seen in Remark 2.6 of the following theorem.

Theorem 2.3. (a) *$\overline{m}_{0,1-\alpha}(\lambda)$ is a conservative $1-\alpha$ UCB for m_0, that is,*

$$\mathrm{pr}(m_0 \le \overline{m}_{0,1-\alpha}(\lambda)) \ge 1-\alpha. \tag{2.13}$$

(b) *Especially, if λ takes the same value as t in the P-value rejection region, then*

$$\mathrm{pr}\Big(Q_t \le \overline{Q}_{1-\alpha}(\overline{m}_{0,1-\alpha}(t), t),\ m_0 \le \overline{m}_{0,1-\alpha}(t)\Big) \ge 1-\alpha. \tag{2.14}$$

Proof. Use $\overline{m}_0$ as a shorthand of $\overline{m}_{0,1-\alpha}(\lambda)$ in this proof. We want to establish that

$$\{m_0 \le \overline{m}_0\} = \{h(m_0) \le h(\overline{m}_0)\}. \tag{2.15}$$

The fact that function $h(k)$ is increasing in k leads to $\{m_0 \le \overline{m}_0\} \subseteq \{h(m_0) \le h(\overline{m}_0)\}$. On the other hand, if $m_0 > \overline{m}_0$, then $\overline{m}_0$ is strictly less than m, and we must have $h(\overline{m}_0) = m - R_\lambda$ according to (2.12). $\overline{m}_0$ is the maximum of k such that $h(k) = m - R_\lambda$, and hence $h(m_0) \neq m - R_\lambda = h(\overline{m}_0)$ as $m_0 > \overline{m}_0$. The increasing property of $h(k)$ leads to $h(m_0) \ge h(\overline{m}_0)$. Combining this with $h(m_0) \neq h(\overline{m}_0)$, we obtain that $h(m_0) > h(\overline{m}_0)$; therefore, we conclude

$$\{m_0 > \overline{m}_0\} \subseteq \{h(m_0) > h(\overline{m}_0)\} \tag{2.16}$$

and complete the proof of (2.15).

Note that

$$\begin{aligned}
\{h(m_0) \le h(\overline{m}_0)\} &= \{h(m_0) \le m - R_\lambda, \overline{m}_0 < m\} \cup \{h(m_0) \le h(m),\ \overline{m}_0 = m\} \\
&= \{h(m_0) \le m - R_\lambda\ , \overline{m}_0 < m\} \cup \{\overline{m}_0 = m\} \quad \big(h(m_0) \le h(m) \text{ always holds}\big) \\
&\supseteq \{h(m_0) \le m - R_\lambda\},
\end{aligned} \tag{2.17}$$

we have

$$\begin{aligned}
\operatorname{pr}(m_0 \le \overline{m}_0) &= \operatorname{pr}\big(h(m_0) \le h(\overline{m}_0) \ge \operatorname{pr}(m_0 - C_{1-\alpha}(m_0, \lambda) \le m - R_\lambda)\big) \\
&\ge \operatorname{pr}(m_0 - C_{1-\alpha}(m_0, \lambda) \le m_0 - V_\lambda) \quad (\text{using } m - R_\lambda \ge m_0 - V_\lambda) \\
&= \operatorname{pr}(V_\lambda \le C_{1-\alpha}(m_0, \lambda)) \\
&\ge 1 - \alpha \quad \Big(\text{Note that } V_\lambda \overset{d}{=} N_{m_0,\lambda}\Big).
\end{aligned} \tag{2.18}$$

Hence, we have the proof of part (a). When λ is set to be t, we have that the set $\{m_0 \le \overline{m}_{0,1-\alpha}(t)\}$ contains $\{V_t \le C_{1-\alpha}(m_0, t)\}$ from the above derivation, and that $\{Q_t \le \overline{Q}_{1-\alpha}(m_0, t)\} = \{V_t \le C_{1-\alpha}(m_0, t)\}$ from the derivation of Lemma 2.2. Therefore,

$$\begin{aligned}
\left\{Q_t \le \overline{Q}_{1-\alpha}(\overline{m}_{0,1-\alpha}(t), t),\ m_0 \le \overline{m}_{0,1-\alpha}(t)\right\} &\supseteq \left\{Q_t \le \overline{Q}_{1-\alpha}(m_0, t),\ m_0 \le \overline{m}_{0,1-\alpha}(t)\right\} \\
&\supseteq \{V_t \le C_{1-\alpha}(m_0, t)\}.
\end{aligned} \tag{2.19}$$

Thus, we have the proof of part (b) of this theorem. □

Remark 2.4. Theorem 2.3 gives researchers a good sense of the total number m_0 of true null hypotheses. Other papers, for example, Storey et al. [10], Benjamini and Hochberg [11], and Langaas et al. [12], gave only point estimates of m_0 or $\pi_0 = m_0/m$. Part (a) gives a confidence inference for m_0, and part (b) gives a simultaneous statement for the Q_t and m_0, which is more interesting. Meinshausen [13] gives a confidence for m_0 by using resampling methods, while ours exploited the independence information so that it works for finite samples.

Remark 2.5. Theorem 2.3 of Göb [14] implies that for a binomial distribution, the difference between the median and mean is less than 1, that is, $|C_{0.5}(m_0, t) - m_0 t| < 1$. From (2.3), we know the 50% UPB for the Q_t can be estimated by $C_{0.5}(m_0, t)/R_t$. This UPB for Q_t is very close to $m_0 t/R_t$ with a difference smaller than $1/R_t$. Replacing m_0 by m in $m_0 t/R_t$ is equivalent to the classical BH procedure. For a very large R_t, the term $1/R_t$ can be ignored, and the BH procedure offers an approximate estimate of the 50% UPB for Q_t.

Remark 2.6. When k is large, the distribution $\mathrm{Bin}(k, \lambda)$ can be closely approximated by $N(k\lambda, k\lambda(1-\lambda))$. Let $z_{1-\alpha}$ be the $1-\alpha$ quantile of a standard normal distribution. After some algebraic manipulations, we obtain a $1-\alpha$ UCB for m_0

$$\overline{m}_{0,1-\alpha}(\lambda) \approx \left\{ \frac{1}{2(1-\lambda)} \left(z_{1-\alpha}\sqrt{\lambda(1-\lambda)} + \sqrt{z_{1-\alpha}^2 \lambda(1-\lambda) + 4(m - R_\lambda)(1-\lambda)} \right) \right\}^2. \tag{2.20}$$

Taking $1-\alpha = 0.5$, we have $(m - R_\lambda)/(1-\lambda)$, which is equivalent to (2.3) of Storey et al. [10]. For most practical applications, one can set the value of $\lambda = 0.5$. Fine tuning of the parameter λ will be discussed in Section 3.3.

When the rejection region $[0, t]$ is small, the UCB for m_0 obtained from part (b) of Theorem 2.3 may be too conservative. It may be advantageous to have separate values for λ and t. Part (a) of Theorem 2.3 implies the following.

Corollary 2.7. *Replacing m_0 by its the upper confidence bound $\overline{m}_{0,1-\alpha_2}(\lambda)$ and α by α_1 in (2.3), we define*

$$\overline{Q}_{1-\alpha_1,1-\alpha_2}(t, \lambda) = \frac{C_{1-\alpha_1}(\overline{m}_{0,1-\alpha_2}(\lambda), t)}{R_t}. \tag{2.21}$$

Then, $\overline{Q}_{1-\alpha_1,1-\alpha_2}(t, \lambda)$ is a conservative $1-\alpha$ $(\alpha = \alpha_1 + \alpha_2)$ UPB for the false discovery proportion Q_t.

3. Upper Prediction Bounds and Simultaneous Inferences

3.1. The Setup

In Section 2, the UPBs for Q are only valid for a *fixed* rejection region $[0, t]$ of P-values. In practice, researchers will not fix the rejected region $[0, t]$ but adapt it to the actual data. The logic is the same as with single hypothesis testing. In single hypothesis testing with nested rejection regions $\{\Gamma_\alpha : 0 < \alpha < 1\}$, for an observed statistic T, one will find the rejection region that contains the observed statistic with the smallest type I error α, that is,

$$P\text{-value}(T) = \min\{\alpha : T \in \Gamma_\alpha\}. \tag{3.1}$$

The same logic can be applied to our false discovery proportion. In this case, we will try to find the largest rejection region $[0, t]$ such that the false discovery proportion Q is not more than γ, say 10%, with probability $1-\alpha$. Define

$$\tau = \max\left\{ t : \overline{Q}_{1-\alpha_1,1-\alpha_2}(t, \lambda) \le \gamma \right\}. \tag{3.2}$$

We then reject any hypothesis whose P-value is no greater than τ. If τ is independent of Q and $\overline{Q}$, then we can expect that

$$\mathrm{pr}(Q_\tau \le \gamma) \ge \mathrm{pr}\left(Q_\tau \le \overline{Q}_{1-\alpha_1,1-\alpha_2}(\tau,\lambda)\right) \ge 1-\alpha. \tag{3.3}$$

Asymptotically, τ and $(Q,\overline{Q})$ may be independent: this question is open for future research. To overcome the independence assumption of τ and $(Q,\overline{Q})$, we seek an alternative approach: to find simultaneous UPBs for all rejection regions $[0,t]$, $t \in (0,1)$, that is, to find an *upper prediction band* $\overline{Q}_t$ such that

$$\mathrm{pr}\left(Q_t \le \overline{Q}_t \text{ for } t \in (0,1)\right) \ge 1-\alpha. \tag{3.4}$$

Hence we have the simultaneous inferences on Q_t for each rejection region $[0,t]$, $t \in (0,1)$. Following the definition of $C_{1-\alpha}(n,t)$ in (2.2) to construct the UPB for Q_t, we want to define the simultaneous critical values of $N_{n,t}$. Using the distribution of $\max_{t\in(0,1)} N_{n,t}$ is unwise as $\max_{t\in(0,1)} N_{n,t} = N_{n,1}$, which takes value n with probability one. A better approach is to center $N_{n,t}$, that is,

$$\sup_{t\in(0,1)} (N_{n,t} - nt). \tag{3.5}$$

This leads to a test statistic related to the Kolmogorov-Smirnov test statistic, which gives an upper confidence band for a cumulative distribution function $F(x)$. It turns out that this method leads to very high UPBs when t is close to zero or one. Therefore, we normalize $N_{n,t}$, that is,

$$\widetilde{Z}_n = \frac{\sup_{t\in(0,1)}(N_{n,t} - nt)}{nt(1-t)}. \tag{3.6}$$

Note that $\widetilde{Z}_n$ is continuously distributed even though each $N_{n,t}$ is discretely distributed. Let $\widetilde{z}_{1-\alpha}(n)$ be the $1-\alpha$ quantile of $\widetilde{Z}_n$, that is,

$$\mathrm{pr}\left(\widetilde{Z}_n \le \widetilde{z}_{1-\alpha}(n)\right) = 1-\alpha. \tag{3.7}$$

We can then redefine $\overline{Q}$ as

$$\widetilde{Q}_{1-\alpha}(m_0,t) = \frac{m_0 t + \widetilde{z}_{1-\alpha}(m_0)\sqrt{m_0 t(1-t)}}{R_t}. \tag{3.8}$$

Corresponding to Lemma 2.2 and Corollary 2.7, we have similar results below.

Corollary 3.1. *For any given* $0 < t < 1$, $\widetilde{Q}_{1-\alpha}(m_0, t)$ *of* (3.8) *is an exactly* $1-\alpha$ *upper prediction band for the false discovery proportion* Q_t, *that is,*

$$\mathrm{pr}\Big(Q_t \le \widetilde{Q}_{1-\alpha}(m_0, t)\ \forall t \in (0,1)\Big) = 1-\alpha. \tag{3.9}$$

Corollary 3.2. *Denote that* $\overline{m}_0 = \overline{m}_{0,1-\alpha_2}(\lambda)$. *Define*

$$\widetilde{Q}_{1-\alpha_1,1-\alpha_2}(t,\lambda) = \frac{\overline{m}_0 t + \widetilde{z}_{1-\alpha_1}(\overline{m}_0)\sqrt{\overline{m}_0 t(1-t)}}{R_t}. \tag{3.10}$$

Let $\alpha = \alpha_1 + \alpha_2$. *Then* $\widetilde{Q}_{1-\alpha_1,1-\alpha_2}(t,\lambda)$ *is a conservative* $1-\alpha$ *upper prediction band for the false discovery proportion* Q_t, *that is,*

$$\mathrm{pr}\Big(Q_t \le \widetilde{Q}_{1-\alpha_1,1-\alpha_2}(t,\lambda)\ \forall t \in (0,1)\Big) \ge 1-\alpha. \tag{3.11}$$

Remark 3.3. Using the same idea as in the proof of Lemma 2.2, the proof of the above corollaries is straightforward after converting the comparison between Q and $\widetilde{Q}$ to the comparison between V_t and $m_0 t + \widetilde{z}_{1-\alpha}(m_0)\sqrt{m_0 t(1-t)}$. This conversion provides a powerful tool for understanding the false discovery proportion.

Remark 3.4. The formulation of $\widetilde{Q}_{1-\alpha}(m_0, t)$ in (3.8) is motivated by the normal approximation of $N_{n,t}$. But our definition of $\widetilde{Q}_{1-\alpha}(m_0, t)$ gives *exact* UPBs simultaneously for all $t \in (0,1)$ due to the exactness of the quantile $\widetilde{z}_{1-\alpha}$.

Remark 3.5. Meinshausen and Rice [6] and Donoho and Jin [15] also utilize the empirical process $\widetilde{Z}_n$. However, they focus on the asymptotic theory for $\widetilde{Z}_n$, which may face the slow convergence problem described in the next section. Our focus is on the finite sample control.

Remark 3.6. Let $f(t) = (k-nt)/\sqrt{nt(1-t)}$. After some simplifications, we have $f'(t) \cdot \left(\sqrt{nt(1-t)}\right)^3 = -n^2 t(1-t) - 1/2 \cdot (k-nt)[n(1-2t)] = -n/2 \cdot [k(1-t) + (n-k)t]$, and then $f(t)$ is a decreasing function in t, $0 < t < 1$ for $0 \le k \le n$. Equation (3.6) can be simplified to be

$$\max_{k=1,\ldots,n} \sup_{t \in [U_{(k)}, U_{(k+1)})} \frac{k-nt}{\sqrt{nt(1-t)}} = \max_{k=1,\ldots,n} \frac{k - nU_{(k)}}{\sqrt{nU_{(k)}\left(1-U_{(k)}\right)}}, \tag{3.12}$$

where $U_{(k)}$ is the kth smallest ordered one among the n samples of $U[0,1]$ distribution. This formulation facilitates the computation of the distribution of $\widetilde{Z}_n$ by Monte Carlo methods. The standard error associated with the Monte Carlo simulations in computing the probability in (3.7) is no greater than $\sqrt{\alpha(1-\alpha)/B}$, where B is the number of simulations.

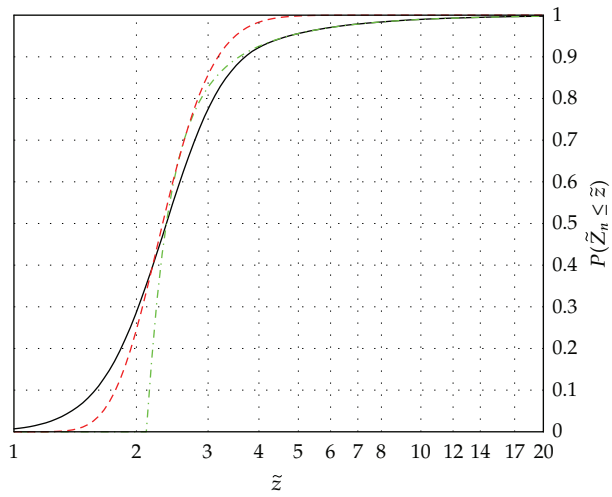

Figure 1: Plot of the probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$ for $1 \leq \tilde{z} \leq 20$ with $n = 10^5$. The x-axis is plotted on a log scale. The black solid curve is computed from 10^6 Monte Carlo simulations. The red-dashed curve is based on (3.13). The green dot-dashed curve is computed from (3.14).

3.2. Computing the Distribution of $\tilde{Z}_n$

In order to make simultaneous inferences, we need to know the distribution of $\tilde{Z}_n$ defined in (3.6). Example 1 of Jaeschke [16] showed that asymptotically, for any x,

$$\lim_{n\to\infty} \text{pr}\left(\tilde{Z}_n \leq \frac{x + 2\ln\ln n + (1/2)\ln\ln\ln n - (1/2)\ln \pi}{\sqrt{2\ln\ln n}}\right) = \exp\left[-\exp(-x)\right]. \tag{3.13}$$

This implies that $\tilde{Z}_n / \sqrt{2\ln\ln n}$ converges to 1 in probability as n goes to ∞. Jaeschke [16] claimed that this probability convergence is of almost no practical use. This is where we need to be cautious using asymptotic results. Figure 1 shows the poor approximation of the asymptotic result, even for a very large $n = 10^5$. Noe and Vandewiele [17] gave an iterative algorithm to compute the exact probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$. Their algorithm is only good for very small n due to the computational time and propagation of precision errors in representing real numbers in computer. Equation (24) of their paper gives an approximate formula for $n = 1, \ldots, 100$,

$$\begin{aligned}\text{pr}\left(\tilde{Z}_n \leq \tilde{z}\right) \approx{}& 1 - (\tilde{z})^{-2} - \left(2 - 3n^{-1}\right)(\tilde{z})^{-4} - \left(10 - 57n^{-1} + 48n^{-2}\right)(\tilde{z})^{-6} \\ &- \left(74 - 1021n^{-1} + 2743n^{-2} - 1797n^{-3}\right)(\tilde{z})^{-8} \\ &- \left(706 - 19123n^{-1} + 111905n^{-2} - 213619n^{-3} + 120132n^{-4}\right)(\tilde{z})^{-10}.\end{aligned} \tag{3.14}$$

This approximation is very good for $\tilde{z} \geq 4$ but is away from the true probability when $\tilde{z} < 4$. For our applications, the 50% quantile (median) of $\tilde{Z}$ is very useful, but the approximation of (3.14) is poor there.

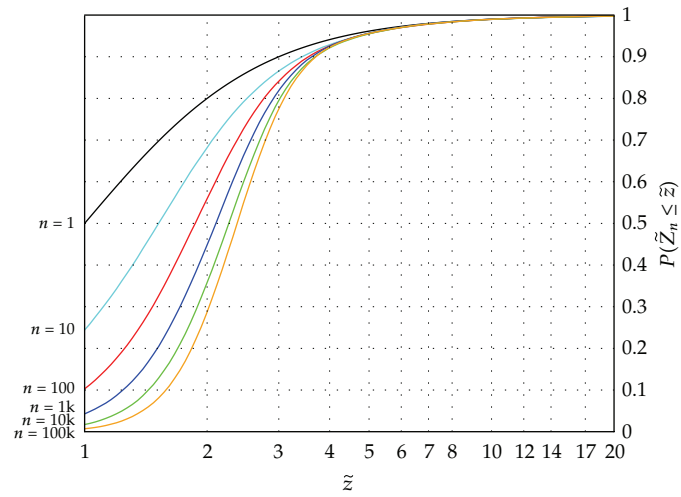

Figure 2: Plot of the probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$ for $1 \leq \tilde{z} \leq 20$. The probability is computed with 10^6 Monte Carlo simulations. The curves from the top to the bottom correspond to $n = 1, 10, 10^2, 10^3, 10^4$, and 10^5.

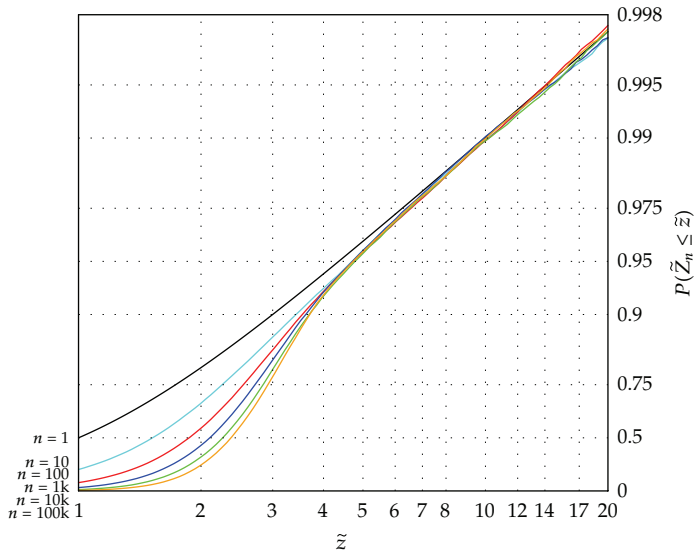

Figure 3: A blowup of Figure 2. This shows the part of the probability $\text{pr}(\tilde{Z}_n \leq \tilde{z})$ that is close to 1 as $1-P$ is drawn on a log scale for the y-axis.

In order to overcome the above poor approximation, we propose to use the Monte Carlo method to obtain the probability $\text{pr}(\tilde{Z} \leq \tilde{z})$. Figures 2 and 3 give the probability for $\tilde{z} \in [1, 20]$ with 10^6 simulations for $n = 1, 10, 10^2, 10^3, 10^4$, and 10^5. The Monto Carlo method generated quantiles $\tilde{z}$ for $n = 1, \ldots, 100$ are almost the same as those quantiles that were able to be computed by the exact algorithm in Noe and Vandewiele [17]. Our two figures show that the distribution of $\tilde{Z}_n$ does not change dramatically from $n = 1, \ldots, 10^5$. This property is beneficial for a multiple testing problem with large number of hypotheses, as it will not be overpenalized. Table 1 gives the quantiles of $\tilde{Z}_n$ with $n = 10^5$.

Table 1: The quantiles of $\widetilde{Z}_n$ of Figure 3, where $n = 10^5$. This table is estimated by 10^6 Monto Carlo simulations. The column $1-\alpha$ gives the probabilities, and the column $\tilde{z}$ gives the quantiles of $\widetilde{Z}_n$.

$1-\alpha$	$\tilde{z}$	$1-\alpha$	$\tilde{z}$	$1-\alpha$	$\tilde{z}$	$1-\alpha$	$\tilde{z}$
0.5	2.37	0.69	2.76	0.88	3.52	0.9675	5.74
0.51	2.39	0.7	2.78	0.89	3.61	0.97	5.95
0.52	2.41	0.71	2.81	0.9	3.7	0.9725	6.19
0.53	2.42	0.72	2.84	0.905	3.75	0.975	6.47
0.54	2.44	0.73	2.86	0.91	3.81	0.9775	6.78
0.55	2.46	0.74	2.89	0.915	3.88	0.98	7.23
0.56	2.48	0.75	2.92	0.92	3.96	0.9825	7.73
0.57	2.5	0.76	2.95	0.925	4.05	0.985	8.31
0.58	2.52	0.77	2.99	0.93	4.15	0.9875	9.04
0.59	2.54	0.78	3.02	0.935	4.26	0.99	10.04
0.6	2.56	0.79	3.06	0.94	4.4	0.9925	11.56
0.61	2.58	0.8	3.09	0.945	4.55	0.995	14.24
0.62	2.6	0.81	3.13	0.95	4.73	0.9955	14.87
0.63	2.62	0.82	3.18	0.9525	4.84	0.996	15.75
0.64	2.64	0.83	3.22	0.955	4.95	0.9965	16.6
0.65	2.66	0.84	3.27	0.9575	5.08	0.997	17.97
0.66	2.69	0.85	3.32	0.96	5.22	0.9975	19.69
0.67	2.71	0.86	3.39	0.9625	5.36		
0.68	2.73	0.87	3.45	0.965	5.55		

3.3. More about the Upper Confidence Bound for m_0

In computing the UCB for m_0 and consequently the UPB for Q_t, we rely on the unspecified parameter λ. A conventional choice of λ is 0.5. It is tempting to use $\min_{\lambda\in(0,1)}\overline{m}_{0,1-\alpha}(\lambda)$ as the best UCB for m_0. This approach should be avoided as it may lead to an overoptimistic UCB. We can use the same idea in computing the simultaneous upper prediction bounds for Q_t to find an UCB for m_0. Equation (2.20) motivates to the following theorem.

Theorem 3.7. *Define* $\widetilde{m}_{0,1-\alpha}(\lambda)$ *as*

$$\left\{\frac{1}{2(1-\lambda)}\left(\overline{\overline{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)}+\sqrt{\left(\overline{\overline{z}}_{1-\alpha}(m)\right)^2\lambda(1-\lambda)+4(m-R_\lambda)(1-\lambda)}\right)\right\}^2, \tag{3.15}$$

where

$$\overline{\overline{z}}_{1-\alpha}(m)=\max_{n=1,\ldots,m}\widetilde{z}_{1-\alpha}(n). \tag{3.16}$$

Let

$$\widetilde{m}_{0,1-\alpha}=\min_{\lambda\in(0,1)}\widetilde{m}_{0,1-\alpha}(\lambda). \tag{3.17}$$

Using $\tilde{m}_{0,1-\alpha}$ to replace m_0 in (3.8) *results in $\tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$. We have*

$$\mathrm{pr}\Big(\mathrm{m}_0 \le \tilde{\mathrm{m}}_{0,1-\mathrm{ff}},\ \mathrm{Q_t} \le \tilde{\mathrm{Q}}_{1-\mathrm{ff}}(\tilde{\mathrm{m}}_{0,1-\mathrm{ff}}, \mathrm{t})\ \forall \mathrm{t} \in (0,1)\Big) \ge 1 - \mathrm{ff}. \tag{3.18}$$

Thus simultaneously $\tilde{m}_{0,1-\alpha}$ is a $1-\alpha$ UCB for m_0 and $\tilde{Q}$ is a $1-\alpha$ upper prediction band.

Proof. Note that when $x > 0$,

$$x(1-\lambda) - \sqrt{x}\bar{\tilde{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)} - (m - R_\lambda) \le 0 \tag{3.19}$$

if and only if

$$x \le \left\{ \frac{1}{2(1-\lambda)} \left(\bar{\tilde{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)} + \sqrt{\left(\bar{\tilde{z}}_{1-\alpha}(m)\right)^2 \lambda(1-\lambda) + 4(m - R_\lambda)(1-\lambda)} \right) \right\}^2. \tag{3.20}$$

Therefore,

$$\begin{aligned}
\{m_0 \le \tilde{m}_{0,1-\alpha}\} &= \{m_0 \le \tilde{m}_{0,1-\alpha}(\lambda)\ \forall \lambda \in (0,1)\} \\
&= \left\{ m_0(1-\lambda) - \sqrt{m_0}\bar{\tilde{z}}_{1-\alpha}(m)\sqrt{\lambda(1-\lambda)} - (m - R_\lambda) \le 0\ \forall \lambda \in (0,1) \right\} \\
&= \left\{ \max_{\lambda \in (0,1)} \frac{m_0(1-\lambda) - (m - R_\lambda)}{\sqrt{m_0 \lambda(1-\lambda)}} \le \bar{\tilde{z}}_{1-\alpha}(m) \right\} \\
&\supseteq \left\{ \max_{\lambda \in (0,1)} \frac{V_\lambda - m_0\lambda}{\sqrt{m_0\lambda(1-\lambda)}} \le \tilde{z}_{1-\alpha}(m_0) \right\}.
\end{aligned} \tag{3.21}$$

The last step follows: (i) $\bar{\tilde{z}}_{1-\alpha}(m)$ is no less than $\tilde{z}_{1-\alpha}(n)$ for any $n \le m$, and (ii) $m - R_\lambda \ge m_0 - V_\lambda$. The fact (ii) gives

$$m_0(1-\lambda) - (m - R_\lambda) \le m_0(1-\lambda) - (m_0 - V_\lambda) = V_\lambda - m_0\lambda. \tag{3.22}$$

Following the same idea as in the proof of Theorem 2.3 part (b), we can show that the set $\{m_0 \le \tilde{m}_{0,1-\alpha}$ and $Q_t \le \tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$ for all $t \in (0,1)\}$ is a superset of $\{\max_{t\in(0,1)}(V_t - m_0 t)/\sqrt{m_0 t(1-t)} \le \tilde{z}_{1-\alpha}(m_0)\}$. Therefore,

$$\begin{aligned}
&\mathrm{pr}\Big(m_0 \le \tilde{m}_{0,1-\alpha},\ Q_t \le \tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)\ \forall t \in (0,1)\Big) \\
&\quad \ge \mathrm{pr}\left(\max_{t\in(0,1)} \frac{V_t - m_0 t}{\sqrt{m_0 t(1-t)}} \le \tilde{z}_{1-\alpha}(m_0) \right) \\
&\quad = 1 - \alpha \quad \Big(\text{Note that } V_t \overset{d}{=} N_{m_0,t}\Big).
\end{aligned} \tag{3.23}$$

For any given α and γ,
(1) Compute $\tilde{m}_{0,1-\alpha}(\lambda)$ of (3.15) for some pre-specified λ_i's, say $\lambda_i = i/1000$, for $i = 1, \ldots, 999$.
(2) Compute $\tilde{m}_{0,1-\alpha} = \min_i \tilde{m}_{0,1-\alpha}(\lambda_i)$. This $\tilde{m}_{0,1-\alpha}$ is the $1-\alpha$ UCB for m_0.
If $\tilde{m}_{0,1-\alpha}$ exceeds m, replace it by m.
(3) Sort the observed P-values such that $P_{(1)} \leq \cdots \leq P_{(m)}$, and use (3.8) to compute the $1-\alpha$ simultaneous UPBs for the false discovery proportion Q, that is, for $i = 1, \ldots, m$,
$\tilde{Q}_{1-\alpha}(P_{(i)}) = (1/i)\ (\tilde{m}_{0,1-\alpha} P_{(i)} + \tilde{z}_{1-\alpha}(\tilde{m}_{0,1-\alpha})\sqrt{\tilde{m}_{0,1-\alpha} P_{(i)}(1 - P_{(i)})})$.
If $\tilde{Q}_{1-\alpha}(P_{(i)})$ exceeds 1, replace it by 1.
(4) Compute $\tau = \max\{P_{(i)} : \tilde{Q}_{1-\alpha}(P_{(i)}) \leq \gamma\}$,
reject the hypotheses whose P-values are no greater than τ,
which ensures that the false discovery proportion Q is not exceeding γ with probability $1-\alpha$.

Algorithm 1: Compute the simultaneous UPBs for the false discovery proportion and the UCB for m_0.

Readers should note that the maximum quantile $\bar{\tilde{z}}_{1-\alpha}(m)$ defined in (3.16) is only used to construct $\tilde{m}_{0,1-\alpha}$, The construction of $\tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$ itself does not use the maximum but the quantile $\tilde{z}_{1-\alpha}(m)$, while $\tilde{Q}_{1-\alpha}(\tilde{m}_{0,1-\alpha}, t)$ still depends on the maximum quantiles indirectly through $\tilde{m}_{0,1-\alpha}$. □

3.4. The Algorithm

Putting all these pieces together, we describe the procedure to compute the upper prediction band for Q_t and the UCB for m_0 in Algorithm 1. Note that we have to compute the quantile $\bar{\tilde{z}}_{1-\alpha}(m)$ and $\tilde{z}_{1-\alpha}(\tilde{m}_{0,1-\alpha})$. This is very time consuming for large m, which is typically from thousands to tens of thousands. The computationally time can be reduced by the following strategies.

(1) After careful study of the two equations (3.8) and (3.15), we find that if we replace all $\tilde{z}_{1-\alpha}(n)$ and $\bar{\tilde{z}}_{1-\alpha}(n)$ by $\bar{\tilde{z}}_{1-\alpha}(N)$, where $N \geq n$, the conclusions of Corollaries 3.1 and 3.2 and Theorem 3.7 still hold.

(2) The quantile $\tilde{z}_{1-\alpha}(n)$ is an increasing function of n, as shown by the Monte Carlo simulations in Figures 2 and 3. The rigorous mathematical proof of this finding is open to future research. For practical applications, we can first use Monte Carlo simulations to verify this property for the range of n that is related to the project and then replace all $\bar{\tilde{z}}_{1-\alpha}(n)$ by $\tilde{z}_{1-\alpha}(n)$.

(3) Figure 2 shows that $\tilde{z}_{1-\alpha}(n)$ is very close to $\tilde{z}_{1-\alpha}(N)$ if n is close to a large N, say more than 100. Therefore, in practical computations, we can first compute and store a representative sequences of the quantiles $\tilde{z}_{1-\alpha}(n)$ for $n = n_1, \ldots, n_I$, and consequently we can get an upper bound for $\tilde{z}_{1-\alpha}(n)$ for $n = 1, \ldots, m$. In computing the quantiles $\tilde{z}_{1-\alpha}(n)$, we recommend to have at least 10^4 Monte Carlo simulations in order to get an accurate quantile computation for tail part. Even with 10^6 simulations, we still see a small amount of random noise in the tail part in Figure 3.

4. A Focused Inference on Q and m_0: A Unified Approach

In many applications, it may be unnecessary to compute the simultaneous UPBs for Q_t for all t in $(0,1)$ and using $\tilde{m}_{1-\alpha}(\lambda)$ for all λ in $(0,1)$ to derive a $1-\alpha$ UCB for m_0. In most applications, it may be reasonable to restrict the rejections onto $\{[0,t] : t \in [t_0, t_0']\}$. The t_0 can take value of $0.01/m$ based on Bonferroni FWER control at level 0.01. It is rare to consider a smaller rejection region than this. The t_0' can take value of 0.05 as it is rare to consider a larger rejection region than $[0, 0.05]$ even in a single hypothesis testing problem. For the same reason, we can also restrict λ onto $[\lambda_0, \lambda_0']$ in (3.17). The interval $[\lambda_0, \lambda_0']$ can be taken as a region close to one as the minimum of $\tilde{m}_{0,1-\alpha}(\lambda)$ is reached when λ is close to 1 [18], but if λ is too close to 1, $\tilde{m}_{0,1-\alpha}(\lambda)$ is not stable. One good choice of λ_0 can be 0.8, and λ_0' can be 0.95.

The above scenario is a focused inference on Q and m_0. The $\tilde{z}$ in Section 3.2 will be a little bit more conservative for us. We can redefine $\widetilde{Z}_n^*$ as in the following:

$$\widetilde{Z}_n^* = \max\left(\frac{\sup_{t\in[t_0,t_0']}(N_{n,t}-nt)}{\sqrt{nt(1-t)}}, \frac{\sup_{\lambda\in[\lambda_0,\lambda_0']}(N_{n,\lambda}-n\lambda)}{\sqrt{n\lambda(1-\lambda)}}\right). \tag{4.1}$$

From this $\widetilde{Z}_n^*$ we can define the $1-\alpha$ quantile $\tilde{z}_{1-\alpha}^*(n)$, and derive results similar to Theorem 3.7. Figure 4 shows quantiles $\tilde{z}_{1-\alpha}^*(n)$ for $n = 1, \ldots, 10^5$ for $[t_0, t_0'] = [0.01/n, 0.05]$ and $[\lambda_0, \lambda_0'] = [0.8, 0.95]$. Table 2 gives the numerical values of $\tilde{z}_{1-\alpha}^*(n)$ for $n = 10^5$. It clearly shows that $\tilde{z}_{1-\alpha}^*(n)$ is around 10% smaller than the unrestricted quantiles $\tilde{z}_{1-\alpha}^*(n)$. For small values of α, say that $\alpha \le 0.01$, the former is at least 25% smaller than the latter.

Corollary 4.1. *Define* $\tilde{m}_{0,1-\alpha}^*(\lambda)$ *as*

$$\left\{\frac{1}{2(1-\lambda)}\left(\overline{\tilde{z}}_{1-\alpha}^*(m)\sqrt{\lambda(1-\lambda)} + \sqrt{\left(\overline{\tilde{z}}_{1-\alpha}^*(m)\right)^2\lambda(1-\lambda) + 4(m-R_\lambda)(1-\lambda)}\right)\right\}^2, \tag{4.2}$$

where $\overline{\tilde{z}}_{1-\alpha}^*(m) = \max_{n=1}^m \tilde{z}_{1-\alpha}^*(n)$. *Let*

$$\tilde{m}_{0,1-\alpha}^* = \min\left(\min_{\lambda\in[\lambda_0,\lambda_0']}\tilde{m}_{0,1-\alpha}^*(\lambda), \min_{\lambda\in[t_0,t_0']}\tilde{m}_{0,1-\alpha}^*(\lambda)\right). \tag{4.3}$$

Define $\widetilde{Q}^*$

$$\widetilde{Q}_{1-\alpha}^*(m_0,t) = \frac{m_0 t + \tilde{z}_{1-\alpha}^*(m_0)\sqrt{m_0 t(1-t)}}{R_t}. \tag{4.4}$$

Replacing m_0 *by* $\tilde{m}_{0,1-\alpha}^*$ *results in* $\widetilde{Q}_{1-\alpha}^*(\tilde{m}_{0,1-\alpha}^*, t)$. *We have that* $\tilde{m}_{0,1-\alpha}^*$ *is a* $1-\alpha$ *UCB for* m_0, *and* $\widetilde{Q}^*$ *is a* $1-\alpha$ *upper prediction band for* Q *for* $t \in [t_0 \cdot t_0']$, *that is,*

$$\mathrm{pr}\left(m_0 \le \tilde{m}_{0,1-\alpha}^*,\ Q_t \le \widetilde{Q}_{1-\alpha}^*\left(\tilde{m}_{0,1-\alpha}^*, t\right)\ \forall t \in [t_0, t_0']\right) \ge 1-\alpha. \tag{4.5}$$

Table 2: The quantiles of $\widetilde{Z}_n^*$ of Figure 4, where $n = 10^5$. The column $1-\alpha$ gives the probabilities, and the column $\widetilde{z}^*$ gives the quantiles of $\widetilde{Z}_n^*$.

$1-\alpha$	$\widetilde{z}^*$	$1-\alpha$	$\widetilde{z}^*$	$1-\alpha$	$\widetilde{z}^*$	$1-\alpha$	$\widetilde{z}^*$
0.5	2.1	0.69	2.55	0.88	3.38	0.9675	5.15
0.51	2.12	0.7	2.58	0.89	3.47	0.97	5.28
0.52	2.14	0.71	2.61	0.9	3.57	0.9725	5.43
0.53	2.16	0.72	2.64	0.905	3.62	0.975	5.58
0.54	2.18	0.73	2.67	0.91	3.68	0.9775	5.78
0.55	2.21	0.74	2.7	0.915	3.74	0.98	5.99
0.56	2.23	0.75	2.73	0.92	3.82	0.9825	6.2
0.57	2.25	0.76	2.77	0.925	3.89	0.985	6.48
0.58	2.27	0.77	2.8	0.93	3.97	0.9875	6.82
0.59	2.3	0.78	2.84	0.935	4.06	0.99	7.23
0.6	2.32	0.79	2.88	0.94	4.18	0.9925	7.73
0.61	2.35	0.8	2.92	0.945	4.3	0.995	8.37
0.62	2.37	0.81	2.96	0.95	4.43	0.9955	8.52
0.63	2.39	0.82	3.01	0.9525	4.52	0.996	8.63
0.64	2.42	0.83	3.06	0.955	4.61	0.9965	8.77
0.65	2.44	0.84	3.11	0.9575	4.7	0.997	8.94
0.66	2.47	0.85	3.17	0.96	4.79	0.9975	9.12
0.67	2.5	0.86	3.23	0.9625	4.89		
0.68	2.52	0.87	3.3	0.965	5.02		

Note that the $1-\alpha$ UCB for m_0 takes not only the minimum of $\tilde{m}^*_{0,1-\alpha}(\lambda)$ for $\lambda \in [\lambda_0, \lambda_0']$, but also the minimum of $\tilde{m}^*_{0,1-\alpha}(t)$ for $t \in [t_0, t_0']$. This advantage is due to the construction of the $\widetilde{Z}_n^*$, which takes maximum over these two intervals. The details of the calculation are summarized in Algorithm 2. The proof of this corollary is the same as that in Theorem 3.7.

By setting $\lambda_0 = \lambda_0' = t$ and $t_0 = t_0' = t$, this corollary is equivalent to Theorem 2.3 through some algebra manipulations, while Theorem 2.3 uses the exact confidence bound from the binomial distribution without relying on the quantiles of $\widetilde{Z}_n^*$. Furthermore, Theorem 3.7 is exact a special case of Corollary 4.1 by setting $\lambda_0 = 0$, $\lambda_0' = 1$, and $t_0 = 0$, $t_0' = 1$ and by considering open intervals rather close intervals. The focused inference thus unifies both the fixed rejection approach and simultaneous approach. We should be cautious of selecting $[t_0, t_0']$ and $[\lambda_0, \lambda_0']$ based on the observed data, which may result in overoptimistic false discovery proportions. These settings have to be decided before the data are generated. A careful study of choosing appropriate values for $[t_0, t_0']$ and $[\lambda_0, \lambda_0']$ is open for future research.

5. Generalizing the Results to Less-Independent Situations

The results of Sections 2, 3, and 4 are based on the assumption that the true null P-values are independently distributed as $U[0,1]$. Given this, we need no further assumptions concerning the false null P-values. This independence assumption can be weakened as in the following:

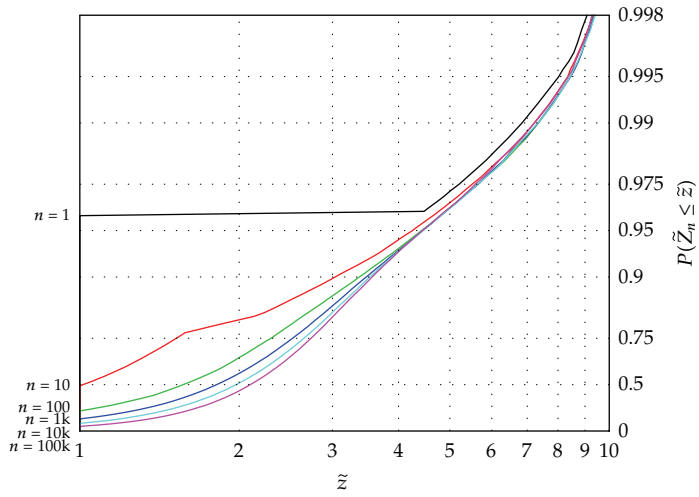

Figure 4: The probability distribution of $\tilde{Z}_n^*$ is computed with 10^6 Monte Carlo simulations. The interval $[t_0, t_0']$ takes value of $[0.01/n, 0.05]$, and interval $[\lambda_0, \lambda_0']$ takes value of $[0.80, 0.95]$. The curves from the top to the bottom correspond to $n = 1, 10, 10^2, 10^3, 10^4$, and 10^5.

For any given α and γ, choose $t_0 = 0.01/m$, $t_0' = 0.05$, $\lambda_0 = 0.8$, $\lambda_0' = 0.95$.
(1) Compute $\tilde{m}^*_{0,1-\alpha}(\lambda)$ of (4.2) for some pre-specified λ_i's in the region (λ_0, λ_0') and pre-specified t_i's in the region (t_0, t_0'), say $\lambda_i = \lambda_0 + (\lambda_0' - \lambda_0)i/1000$, $t_i = t_0 + (t_0' - t_0)i/1000$ for $i = 0, \ldots, 1000$.
(2) Compute $\tilde{m}^*_{0,1-\alpha} = \min(\min_i \tilde{m}^*_{0,1-\alpha}(\lambda_i), \min_i \tilde{m}^*_{0,1-\alpha}(t_i))$. This $\tilde{m}^*_{0,1-\alpha}$ is the $1-\alpha$ UCB for m_0. If $\tilde{m}^*_{0,1-\alpha}$ exceeds m, replace it by m.
(3) Sort the observed P-values such that $P_{(1)} \leq \cdots \leq P_{(m)}$, and use (4.4) to compute the $1-\alpha$ UPB for the false discovery proportion Q, that is, for $P_{(i)} \in [t_0, t_0']$
$\tilde{Q}^*_{1-\alpha}(P_{(i)}) = (1/i)(\tilde{m}^*_{0,1-\alpha} P_{(i)} + \tilde{z}^*_{1-\alpha}(\tilde{m}^*_{0,1-\alpha})\sqrt{\tilde{m}^*_{0,1-\alpha} P_{(i)}(1 - P_{(i)})})$.
If $\tilde{Q}^*_{1-\alpha}(P_{(i)})$ exceeds 1, replace it by 1.
(4) Compute $\tau = \max\{P_{(i)} \in [t_0, t_0'] : \tilde{Q}^*_{1-\alpha}(P_{(i)}) \leq \gamma\}$, reject the hypotheses whose P-values are no greater than τ, which ensures that the false discovery proportion Q is not exceeding γ with probability $1 - \alpha$.

Algorithm 2: Focused simultaneous inferences on the UPBs for the false discovery proportion and the UCB for m_0.

Binomial Dominant Condition: One has $V_t \overset{d}{\leq} N_{m_0,t}$ for $0 < t < 1$.

The notation $X \overset{d}{\leq} Y$ means that random variable X is stochastically no greater than random variable Y, that is, $\text{pr}(X \leq x) \geq \text{pr}(Y \leq x)$ for any x. Replacing the independence assumption by the binomial dominant condition, the results corresponding to Lemma 2.2, Theorem 2.3, and Corollary 2.7 in Section 2 still hold for a fixed rejection region. For the simultaneous UPBs in Sections 3 and 4, we need a stronger assumption than the binomial dominant condition as the joint distribution of $\{V_t, t \in (0,1)\}$ needs to be specified. We can replace the binomial dominant condition by the following.

Joint Binomial Dominant Condition: $(V_{t_1}, \ldots, V_{t_k}) \overset{d}{\leq} (N_{m_0,t_1}, \ldots, N_{m_0,t_k})$ for any $k = 1, 2, \ldots$, and $t_1, \ldots, t_k \in (0,1)$. Here $N_{m_0,t} = \sum_{i=1}^{m_0} I(U_i \leq t)$, and U_i, $i = 1, \ldots, m_0$ are mutually independently distributed as distribution $U[0,1]$. The notation $(X_1, \ldots, X_k) \overset{d}{\leq} (Y_1, \ldots, Y_k)$ means for any $x_1, \ldots, x_k$, $\text{pr}(X_1 \leq x_1, \ldots, X_k \leq x_k) \geq \text{pr}(Y_1 \leq x_1, \ldots, Y_k \leq x_k)$.

Replacing the independence assumption by this joint binomial dominant condition, the results in Sections 3 and 4 are still valid for the upper prediction band for Q and the UCB for m_0. A special case for this joint binomial dominant condition is that when the true null P-values are independent with distribution stochastically no smaller than $U[0,1]$. This happens when the null hypothesis is composite or the statistic to test the null hypothesis is not a continuous random variable.

More generally, we would like the construction of upper prediction band for Q not to rely on the independence assumption or any kind of weak dependence assumption (the binomial dominant condition or the joint binomial dominant condition). The method of Romano and Shaikh [7, 8] can be applied without any assumptions on the dependence, but may potentially have lost power due to that the correlation structure of the data has not been exploited. A resampling procedure [13, 19] has been proposed to address this limitation.

6. Discussion

The method of this paper applies to data where true null P-values are independent, or to slightly dependent data where the joint binomial dominant condition is satisfied. This assumption does not rely on any specification for the false null P-values. In this paper we used the idea of considering a fixed rejection region to construct a UPB for Q_t and a UCB for m_0. By utilizing the normalized empirical process $\widetilde{Z}_n = \sup_{t\in(0,1)}(N_{n,t} - nt)/\sqrt{nt(1-t)}$, we find simultaneous UPBs for Q_t for all $t \in (0,1)$ and can further modify the construction of the UCB for m_0. The result of Theorem 3.7 gives the joint statement about the UCB for m_0 and the simultaneous UPBs for the false discovery proportions Q. A focused approach in Corollary 4.1 unifies the result of the fixed rejection region method and the simultaneous approach.

The method in this paper is based on finite samples and avoids the slow convergence problem of the asymptotic theory for the empirical process $\widetilde{Z}_n$. The Monte Carlo simulations give very accurate estimates of the quantiles for $\widetilde{Z}_n$. The standard error associated with the Monte Carlo simulations in computing the probability in (3.7) is no greater than $\sqrt{\alpha(1-\alpha)/B}$, where B is the number of simulations.

In the dataset where the test statistics are not independent or do not satisfy joint binomial dominant condition, the method in this paper may not be guaranteed to work. One can alternatively use the methods proposed in Romano and Shaikh [7, 8], Meinshausen [13], Ge et al. [19]. The method proposed in this paper can be potentially extended to dependent data by using resamplings, and this work is open for future research.

Acknowledgments

The authors thank Terry Speed, Stuart Sealfon, Carol Bodian, Sylvan Wallenstein, John Mandeli, Samprit Chatterjee, and Jim Godbold for their discussions of this work. They thank the editor and referees for helpful comments that have led to an improved paper. This

work was partly supported by National Institute of Allergy and Infectious Diseases with the contract HHSN266200500021C.

References

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.
[2] C. Genovese and L. Wasserman, "Operating characteristics and extensions of the false discovery rate procedure," *Journal of the Royal Statistical Society B*, vol. 64, no. 3, pp. 499–517, 2002.
[3] E. L. Korn, J. F. Troendle, L. M. McShane, and R. Simon, "Controlling the number of false discoveries: application to high-dimensional genomic data," *Journal of Statistical Planning and Inference*, vol. 124, no. 2, pp. 379–398, 2004.
[4] C. Genovese and L. Wasserman, "A stochastic process approach to false discovery control," *The Annals of Statistics*, vol. 32, no. 3, pp. 1035–1061, 2004.
[5] E. L. Lehmann and J. P. Romano, "Generalizations of the familywise error rate," *The Annals of Statistics*, vol. 33, no. 3, pp. 1138–1154, 2005.
[6] N. Meinshausen and J. Rice, "Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses," *The Annals of Statistics*, vol. 34, no. 1, pp. 373–393, 2006.
[7] J. P. Romano and A. M. Shaikh, "On stepdown control of the false discovery proportion," in *Lehmann Symposium—Optimality*, vol. 49 of *IMS Lecture Notes—Monograph Series*, pp. 33–50, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2nd edition, 2006.
[8] J. P. Romano and A. M. Shaikh, "Stepup procedures for control of generalizations of the familywise error rate," *The Annals of Statistics*, vol. 34, no. 4, pp. 1850–1873, 2006.
[9] M. J. van der Laan, S. Dudoit, and K. S. Pollard, "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*, vol. 3, no. 1, article 15, 2004.
[10] J. D. Storey, J. E. Taylor, and D. Siegmund, "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach," *Journal of the Royal Statistical Society B*, vol. 66, no. 1, pp. 187–205, 2004.
[11] Y. Benjamini and Y. Hochberg, "On the adaptive control of the false discovery rate in multiple testing with independent statistics," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 1, pp. 60–83, 2000.
[12] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to DNA microarray data," *Journal of the Royal Statistical Society B*, vol. 67, no. 4, pp. 555–572, 2005.
[13] N. Meinshausen, "False discovery control for multiple tests of association under general dependence," *Scandinavian Journal of Statistics*, vol. 33, no. 2, pp. 227–237, 2006.
[14] R. Göb, "Bounds for median and 50 percentage point of binomial and negative binomial distribution," *Metrika*, vol. 41, no. 1, pp. 43–54, 1994.
[15] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *The Annals of Statistics*, vol. 32, no. 3, pp. 962–994, 2004.
[16] D. Jaeschke, "The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals," *The Annals of Statistics*, vol. 7, no. 1, pp. 108–115, 1979.
[17] M. Noe and G. Vandewiele, "The calculation of distributions of Kolmogorov-Smirnov type statistics including a table of significance points for a particular case," *Annals of Mathematical Statistics*, vol. 39, pp. 233–241, 1968.
[18] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.
[19] Y. Ge, S. C. Sealfon, and T. P. Speed, "Multiple testing and its applications to microarrays," *Statistical Methods in Medical Research*, vol. 18, no. 6, pp. 543–563, 2009.