A comprehensive error rate for multiple testing

Djalel Eddine Meskaldji
Jean-Philippe Thiran
LTS5/STI, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.
Stephan Morgenthaler
FSB/MATHA, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

Summary. In multiple testing, a variety of control metrics have been introduced such as the Family-Wise Error Rate (FWER), the False Discovery Rate (FDR), the False Exceedance Rate (FER), etc. We found a way to embed these metrics into a continuous family of control metrics, all of which can be attained by applying a simple and general family of multiple testing procedures. The new general error rate (GER) limits the number of false positives relative to an arbitrary non-decreasing function of the number of rejections. An example is $V/R^\beta$, the number of false rejections divided by the number of rejections to a power $0 \leq \beta \leq 1$. We investigated both the control of quantiles and of expectations and provide the corresponding multiple testing procedures. In the above example, the expectation of the criterion thus leads to a family of multiple testing procedures that bridges the gap between the FWER and the FDR.

Keywords: Multiple comparisons, Family-Wise Error Rate, False Discovery Rate, ordered p-values.

1. Introduction

The problem of multiple testing is a key technology in a variety of modern applications such as genomics or neuroimaging. If $m$ hypotheses are tested and if each hypothesis is tested separately at significance level $\alpha$ (without considering multiplicity), the probability of observing at least one false significant result (the FWER), even if there is no real effect and if the tests are independent, is

$$\text{FWER} = P(\text{at least one significant result}) = 1 - (1 - \alpha)^m. \quad (1)$$

For example, if $m = 100$, and $\alpha = 0.05$, this probability is 0.994. The expected number of false positives (the Per Family Error Rate, PFER) is $\alpha m = 5$. Both quantities increase with $m$ and are out of control. Although the evidence that a correction for multiplicity should be mandatory, a large number of claims are published without a proper control. See Benjamini (2010). A general outcome of multiple comparisons is summarized in Table 1. The idea that the control of false positives $V$ should be considered in conjunction with the number of rejections $R$ has been widely accepted by users after the introduction of the False Discovery Proportion (FDP) $\frac{V}{s(R)}$. Nevertheless, in situations where false discoveries have expensive consequences, the FWER remains a viable criterion. We define our general error rate as $\frac{V}{s(R)}$, where $s(R)$ is a scaling function which typically grows more slowly than $R$ itself. This simple device covers and generalizes almost all the existent error rates. In section 4, we

E-mail: djalel.meskaldji@epfl.ch
propose procedures that control the general error rate under different assumptions. The main results are the control of a tail probability by an adaptive step-down (SD) procedure under the Simes (1986) inequality, the control of a tail probability under any assumption, and the control of the expectation under independence using a step-up (SU) procedure.

### 2. Historical background and motivation

Traditional multiple comparisons procedures (MCPs) are designed to control the FWER \( P(V > 0) \), where control at level \( \alpha \) means that FWER \( \leq \alpha \). This is achieved by the Bonferroni (1936) procedure which performs each of the \( m \) tests at level \( \alpha/m \) or equivalently, by rejecting the hypothesis \( H_i \) \( (i = 1, \ldots, m) \) if its corresponding \( p \)-value is less than \( \alpha/m \).

The Bonferroni procedure is the simplest and the strongest procedure in terms of control of \( V \). It even controls the PFER at level \( \alpha \) which is stricter than the FWER. However, when \( m \) grows, the power of the Bonferroni procedure at any fixed alternative tends to 0.

Many other MCPs that control the FWER have been proposed, although they typically give only a slight improvement over the Bonferroni procedure. They all compare the ordered \( p \)-values to thresholds which depend on the global control level \( \alpha \) and the rank of the \( p \)-value. A SU procedure compares the ordered \( p \)-values with the critical thresholds beginning with the larger and thus the less significant \( p \)-value and starts to reject hypothesis once a \( p \)-value is less than its corresponding threshold. After this first crossing, it rejects all hypotheses with smaller \( p \)-values. A SD procedure begins the comparison starting with the smallest or the most significant \( p \)-value. If it is greater than its corresponding threshold then, no hypothesis is rejected. Otherwise, reject hypotheses as long as their \( p \)-values are less than their corresponding thresholds and stop rejecting once a \( p \)-value exceeds its corresponding threshold. Examples of step-wise procedures include Holm (1979), Simes (1986), Hochberg (1988). Note that a SU procedure has a power that equals or exceed the power of a SD procedure that uses the same critical thresholds. See Horn and Dunnett (2004).

Safeguards against false positives is not the unique purpose of testing. Detecting real effects is also of great importance. Benjamini and Hochberg (1995) introduced the FDR as an alternative to the FWER with the aim of increasing power. The FDR is defined to be the expected value of the FDP. We have \( \text{FDP} = V/R \) if \( R > 0 \) and otherwise \( \text{FDP} = 0 \), while the FDR is defined as \( \text{FDR} = E(\text{FDP}) = E \left[ \frac{V}{R} \right| R > 0 \right] P(R > 0) \). Since \( \text{FDR} \leq P(V > 0) \), the FDR is less stringent than the FWER, which should lead to higher power. The FDR has the same behavior as the FWER when all hypotheses are true which implies \( V = R \). Thus, the FWER is weakly controlled. Using an FDR controlling procedure, the number of false positives increases with the number of rejection \( R \). Despite this drawback, the FDR has been widely adopted in many fields of application and it is fair to say that the paper of Benjamini and Hochberg (1995) had a huge impact on the practice of statistics. It has been cited more than 10'000 times up to now. Alternatives to the FDR are available.
Table 2. Summary of some existing error metrics.

| Tail probability | Expectation |
|------------------|-------------|
| FWER = $P(V \geq 0) \leq \alpha$ | PFER = $E(V) \leq \alpha$ |
| k-FWER = $P(V \geq k) \leq \alpha$ | PCER = $E(V/m) \leq \alpha$ |
| Dependent on R   | FDR = $E(V/R) \leq \gamma$ |
| FER = $P(V/R > \gamma) \leq \alpha$ | k-FDR = $E((V-k)_+/R) \leq \gamma$ |
| pFDR = $E(V/R| R > 0) \leq \gamma$ |

Vector (1982), for example, considered the k-FWER = $P(V \geq k)$, which tolerates more false positives and thus increases the power. This seems appropriate when the number of hypotheses $m$ is large. Hommel and Hoffmann (1988) and Lehmann and Romano (2005) derived a single step and a step-down procedures to control the k-FWER. The single step procedure is identical to the Bonferroni procedure except that the p-values are compared to $k\alpha/m$ instead of $\alpha/m$. This procedure is evidently more powerful than the Bonferroni procedure. However, the weak control of the FWER at level $\alpha$ is no longer guaranteed. In fact, the expected number of false positives under the complete null hypothesis ($m_0 = m$) is $k\alpha$. Lehmann and Romano (2005) derived step-wise procedures to control another alternative error metric, the FER, which is defined by $P(FDP > \gamma)$ with $\gamma \in (0,1)$. Many other concepts of false positives error rates have been proposed in the literature. All these concepts have a certain control of false positives situated in between two extremes, the Per Comparison Error Rate (PCER) and the PFER control. Dudoit and van der Laan (2008) and Benjamini (2010) are good sources for additional information.

The false positives metrics can be grouped by two important criteria. First, one can distinguish between metrics that control the probability of exceeding a constant such as the FWER, the k-FWER or the FER and metrics that control the expected number of a certain quantity such as the PFER, the FDR, the k-FDR or the positive FDR (pFDR) (Storey (2002)). Second, one distinguishes between metrics that do not consider the number of rejections such as the FWER, the k-FWER or the PFER and metrics that tolerate more false positives as more hypotheses are rejected, such as the FDR, the pFDR or the FER. Table 2 summarizes this information.

The error rates that control the proportion of false positives are especially appealing for large scale testing problems, compared to error rates that do not consider the number of rejections, as they remain stable when the number of tests $m$ increases. See Dudoit and van der Laan (2008).

The range of metrics is confusing, especially for non experts. In addition, most of the procedures need additional assumptions in order to become attainable. If one or more of theses assumptions are not satisfied, a control failure results. According to Benjamini (2010), none of the metrics is superior in all aspects. In fact users may wish to have several distinct controls achieved by one procedure. For example, one might be willing to derive a powerful procedure that weakly controls the FWER. FDR control procedures seems ideal, but at the risk of large values of $V$ when $R$ becomes large. To avoid this, one might want to add strong control of the k-FWER as the single step of Hommel and Hoffmann (1988) does. The general error rate offers such compromises. If we choose $s_k(R) = R$ for $R \leq k$ and $s_k(R) = k$ for $R \geq k$, the general error rate demands for control of $V/R$ for up to a certain number of rejections, but then switches to a fixed control of $V/k$ thereafter (cut-off FDR).

For $k = 1$, this results in control of the FWER, while for $k = \infty$ we obtain the FDR. This is another example of a family of scaling functions that bridges the gap between these two.
D. E. Meskaldji, J.-Ph. Thiran and S. Morgenthaler

extremes. We can easily show that for independent or positively dependent test statistics (Benjamini and Yekutieli (2001)), the following procedure controls the FDR to be less than \( \gamma \) and controls the PFER to be less than \( k\gamma \) which implies the control of the \( k \)-FWER at level \( \gamma \) since \( P( V \geq k ) \leq E \left( \frac{V}{E} \right) \) by Markov’s inequality.

**Procedure 2.1.** We test \( m \) hypotheses. Let \( p^{(1)} \leq p^{(2)} \leq \cdots \leq p^{(m)} \) be the ordered p-values, and denote by \( H^{(i)} \) the null hypothesis that corresponds to \( p^{(i)} \). Let \( \tilde{i} \) be the largest \( i \) that satisfies \( p^{(i)} \leq \frac{i\gamma}{m} \) as well as \( p^{(i)} \leq \frac{k\gamma}{m} \), that is, \( p^{(i)} \leq \frac{s_{k(i)}\gamma}{m} \); then reject all \( H^{(i)} \) with \( i = 1, \ldots, \tilde{i} \).

Theorem 4.6 gives a general result of this type. The use of such a procedure would be appropriate, particularly in fields where the number of hypotheses tested \( m \) is very large (of the order of \( 10^5 \) or even \( 10^6 \)). Examples are fMRI as well as some genomic studies. This procedure has the control of the FDR when the number of rejections \( R \) is small, but limits the expected number of false positives as \( R \) grows.

For illustration, consider a case, where \( m = 10^4 \) independent tests are performed whose distribution under the null is \( \mathcal{N}(0,1) \). Among the \( m \) tests, \( m_1 = 5 \times 10^2 \) correspond to false hypotheses, in which case the distribution of the test statistics is \( \mathcal{N}(\Delta,1) \), with \( \Delta = 2 \). By setting \( \gamma = 0.05 \) and \( k\gamma = 1 \), the power of the FDR and the cut-off FDR are 0.0505 and 0.0413, and the expected number of false positives \( E(V) \) are 1.359 and 0.881. When the effect \( \Delta \) grows to 4, the respective powers are 0.879 and 0.611 and the expected numbers of false positives are 21.791 and 0.991. Clearly, the FDR achieves its “superior” power by also rejecting a high number of true hypotheses. In many situations, having around 20 false positives is completely out of the question. This example shows that the cut-off FDR gives an added protection often without a high cost in term of power. Note that the theoretic value of the power of the Bonferroni procedure in this situation is 0.00783 and 0.3383 for \( \Delta = 2 \) and \( \Delta = 4 \) respectively.

### 3. The general error rate

The FDP takes into account the number of rejections when controlling false positives. This concept has a Bayesian background in the sense that \( R \) contains information about \( m_1 \), which should be exploited. Our general error rate provides a generalization of the FDP.

**Definition 3.1.** The scaled or general False Discovery Proportion \( s\text{FDP} \) with parameter \( l \geq 1 \) and non-decreasing scaling function \( s : \{1, \ldots, m\} \rightarrow (0, \infty) \) is defined as

\[
s\text{FDP}_l = \begin{cases} 
\frac{V}{s(R)} & \text{if } V \geq l, \\
0 & \text{otherwise}.
\end{cases}
\]

Based on this quantity we define two types of general error rates using two different stochastic functions. The first uses the tail probability of exceeding a constant, which we denote as scaled or general Tail Probability (\( s\text{TP} \)). The second uses the expected value of the \( s\text{FDP} \), which we denote as scaled or general expected value (\( s\text{EV} \)). Remotely related concepts were introduced in van der Laan et al. (2004) and described in Dudoit and van der Laan (2008). These authors consider transformations which involves both \( V \) and \( R \), while we concentrate on the denominator. The reason for doing so can be gleaned from

\[
\frac{V}{s(R)} = \frac{V}{R} \cdot \frac{R}{s(R)},
\]

(2)
A comprehensive error rate

which shows that control of the sFDP\(_1\) (sFDP from now on) is equivalent to control of the FDP times a positive multiplier that depends on \(R\) and that could be greater or less than 1 depending on the level of conservativeness that the researcher desires. In addition, when the scaling function \(s\) is a constant, the sFDP depends only on \(V\). This means that in this case, the relation between the number of false positives \(V\) and the number of rejections \(R\) is suppressed.

3.1. The scaled Tail Probability (sTP) error rate

Many error rates introduced in the literature are tail probabilities of exceeding a certain threshold such as the FWER, k-FWER and the FER. We define the sTP\(_\gamma\) as the probability that the sFDP exceeds a non-negative constant \(\gamma \geq 0\),

\[
sTP_{\gamma} = P[sFDP > \gamma].
\]

sTP\(_\gamma\) with \(s(R) = R\) and \(\gamma \in (0, 1)\) is identical to the FER, while sTP\(_\gamma\) with \(s(R) \cdot \gamma = (k - 1)\) is identical to the k-FWER. In addition, sTP\(_\gamma\) with \(s(R) > 0\) and \(\gamma = 0\) becomes \(P(V > 0)\), the FWER.

The control of the sTP\(_\gamma\) implies the control of quantiles of the sFDP, because sTP\(_\gamma\) \(\leq \alpha\) implies that the \(1 - \alpha\) quantile of sFDP is smaller than \(\gamma\). In particular, when \(\alpha = 0.5\) sTP\(_\gamma\) \(\leq \alpha\) is equivalent to \(\text{median}(sFDP) \leq \gamma\).

3.2. The scaled Expected Value (sEV) error rate

To comprehend some of the false positives error rates which are defined as the expected number of a certain random quantity, such as the PFER and the FDR, we define the sEV by

\[
sEV = E[sFDP].
\]

When \(l = 1\), sEV with \(s(R) = R\) is identical to the FDR of Benjamini and Hochberg (1995), while for \(l = k \geq 1\), and \(s(R) = R\), it is identical to the k-FDR of Sarkar (2007). In addition, sEV with \(s(R) = 1\) and \(l = 1\) becomes \(E(V)\), which is the PFER.

One can also generalize some other concepts in the same way. For example, we define the positive sEV by

\[
E\left[\frac{V}{s(R)} \Bigr| R > 0\right].
\]

Note that for both sTP and sEV, when the scaling function is a constant, we find the error rates with names containing the word "family". Furthermore, if \(m = m_0\),

\[
sEV = E\left[\frac{R}{s(R)} \Bigr| R > 0\right] P[R > 0] = E\left[\frac{R}{s(R)} \Bigr| R > 0\right] \times \text{FWER}. \quad (3)
\]

This shows that if \(s(R) \leq R\) for any \(R\) in \(1, \ldots, m\), the control of the sEV at level \(\gamma\) implies the weak control of the FWER at level \(\gamma\).
4. Control procedures

In this section we present generalizations of some existing procedures to control the sTP and the sEV. The procedures proposed are not limited to particular choices but for any choice of the scaling function. We show that some of the existing procedures can be simply modified to give a more general control.

4.1. Procedures that control the sTP

4.1.1. The generalized Lehmann and Romano procedure

Lehmann and Romano (2005) proposed a SD procedure to control the FER. Here, we give a simply modified version of their procedure to control the sTP.

**Procedure 4.1.** Let \( p(1) \leq p(2) \leq \cdots \leq p(m) \) be the ordered p-values of \( m \) tests, and denote by \( H(i) \) the null hypothesis that corresponds to \( p(i) \). Set

\[
\alpha_i = \begin{cases} 
\left\lfloor \gamma s(i) \right\rfloor + 1, 
& \text{if } i \leq \left\lfloor \gamma s(i) \right\rfloor + 1 \\
\frac{m}{\gamma s(i) + 1 - i}, & \text{if } i > \left\lfloor \gamma s(i) \right\rfloor + 1
\end{cases}
\]

(4)

If \( p(1) > \alpha_1 \), then reject no hypothesis; otherwise, reject all hypotheses \( H(1), \ldots, H(i) \), where \( i \) is the largest index satisfying

\[
p(1) \leq \alpha_1, \ldots, p(i) \leq \alpha(i).
\]

(5)

Note that if \( s(i) = i \) and \( 0 < \gamma < 1 \), this SD procedure is equal to the procedure proposed by Lehmann and Romano (2005) (LR05 from now on) for controlling the FER. Furthermore, if \( s(i) \) is a constant and \( \gamma s(i) \equiv k - 1 \), we find the critical values of the SD procedure of Lehmann and Romano (2005) to control the k-FWER and of course, if \( \gamma = 0 \), we find the Holm procedure that controls the FWER.

We already know by Lehmann and Romano (2005) that for the case where \( \gamma s(i) \) is constant, the procedure defined above controls the sTP\( _\gamma \) at level \( \alpha \), under any dependency assumption of the p-values. The following theorem states the control in the case where \( \gamma s(i) \) is not a constant.

**Theorem 4.2.** Denote by \( q(1) \leq \cdots \leq q(m_0) \) the ordered p-values corresponding to the \( m_0 \) true null hypotheses. Set \( M = \min \{ \gamma s(m) + 1, m_0 \} \).

(i) For the step-down procedure with \( \alpha_i \) defined in procedure 4.1, we have

\[
P[sFDP > \gamma] \leq P \left[ \bigcup_{k=\gamma s(1)+1}^{M} \left\{ q(k) \leq \frac{k\alpha}{m_0} \right\} \right].
\]

(6)

(ii) Therefore, if the joint distribution of the p-values corresponding to the null hypotheses satisfies the Simes inequality, that is

\[
P \left[ \bigcup \left\{ q(1) \leq \frac{\alpha}{m_0} \right\} \cup \left\{ q(2) \leq \frac{2\alpha}{m_0} \right\} \cup \ldots \cup \left\{ q(m_0) \leq \frac{m_0\alpha}{m_0} \right\} \right] \leq \alpha,
\]

(7)

then \( P[sFDP > \gamma] \leq \alpha. \)
The Simes inequality holds for many joint distributions of positively dependent variables. Sarkar (1998) showed that the Simes inequality holds for any multivariate positive distributions of order 2 (MTP2). Obviously, the condition (6) is less strict than the Simes inequality condition. For the particular case where \(|\gamma s(i)| + 1 = c\) (a constant), the right side of (6) holds for any dependency distribution of the p-values. To show this, note that

\[
P \left[ \bigcup_{k=[\gamma s(1)]+1}^{M} \left\{ q(k) \leq \frac{k\alpha}{m_0} \right\} \right] = P \left[ \bigcup_{c}^{c} \left\{ q(k) \leq \frac{k\alpha}{m_0} \right\} \right] = P \left[ q(c) \leq \frac{c\alpha}{m_0} \right] \leq \alpha. \tag{8}
\]

In the general case, that is, when \(|\gamma s(i)| + 1\) is not a constant, the following lemma stated by Lehmann and Romano (2005), can be used to give a sharp upper bound for the right side of equation (6).

**Lemma 4.3.** (Lemma 3.1 in Lehmann and Romano (2005)) Let \(p_1, ..., p_n\) be \(n\) p-values that satisfy \(P\{p_i \leq u\} \leq u\) for all \(i = 1, ..., n\) and for any \(u \in (0, 1)\). Let \(0 \leq \beta_1 \leq \beta_2 \leq ... \leq \beta_h \leq 1\) for some \(1 \leq h \leq n\). Then

\[
P \left[ \bigcup_{i=1}^{h} \{p(i) \leq \beta_i\} \right] \leq n \sum_{i=1}^{h} (\beta_i - \beta_{i-1}) \frac{1}{i}.
\]

The previous lemma leads to the the following result.

**Theorem 4.4.** If the critical values \(\alpha_i\) are replaced by

\[
\alpha'_i = \frac{\alpha_i}{C_{([\gamma s(1)]+1, [\gamma s(m)])}}
\]

with \(C_{l,h} = \sum_{i=l}^{h} \frac{1}{i}\), then \(P\{s\text{FDP} > \gamma\} \leq \alpha\) for any dependency of the p-values corresponding to the true null hypotheses.

**Proof.** By replacing in lemma 4.3, \(h\) and \(n\) by \(M\) and \(m_0\) respectively and by setting \(\beta_i = 0\) for \(i = 1, ..., [\gamma s(1)]\) and \(\beta_i = \frac{\alpha}{m_0}\) for \(i = [\gamma s(1)] + 1, ..., M\), we obtain

\[
P \{s\text{FDP} > \gamma\} \leq P \left[ \bigcup_{i=[\gamma s(1)]+1}^{M} \left\{ p(i) \leq \frac{i\alpha}{m_0} \right\} \right] \leq m_0 \sum_{i=[\gamma s(1)]+1}^{M} \left( \frac{\alpha}{m_0} \right) \frac{1}{i} \leq \alpha \sum_{i=[\gamma s(1)]+1}^{M} \frac{1}{i}.
\]

It suffices then to replace \(\alpha\) by \(\frac{\alpha}{C_{([\gamma s(1)]+1, [\gamma s(m)])}}\) to have \(P\{s\text{FDP} > \gamma\} \leq \alpha\).

The constant \(C_{([\gamma s(1)]+1, [\gamma s(m)])}\) is usually greater than 1, which means that the control under any assumption is more strict than under Simes inequality. This constant may be less than one in some particular cases depending on the value of \([\gamma s(1)]+1\) but this could happen only when \(\gamma s(1)\) is greater than 1 which is less frequent. In addition, if the lower index is 1, the constant \(C\) is greater than 1. Depending on the scaling function and the value of \([\gamma s(m)]\), the constant \(C\) could be greater or smaller than the one proposed in Lehmann and Romano (2005).
4.2. Procedures that control the sEV

4.2.1. The generalized Benjamini and Hochberg procedure

The Benjamini and Hochberg (1995) procedure can be simply modified to obtain a procedure that controls the sEV.

Procedure 4.5. Let \( p(1) \leq p(2) \leq \cdots \leq p(m) \) be the ordered p-values of \( m \) tests, and denote \( H(i) \) the null hypothesis that corresponds to \( p(i) \). Let \( \tilde{i} \) be the largest \( i \) that satisfies \( p(i) \leq \frac{s(i)}{m} \gamma \); then reject all \( H(i) \) with \( i = 1, 2, \ldots, \tilde{i} \).

Note that if \( s(i) = i \), the procedure is the same as the procedure proposed by Benjamini and Hochberg (1995) (BH95 from now on) to control the FDR. If \( s(i) \equiv 1 \), the procedure becomes the Bonferroni procedure. Furthermore, if \( s(i) = k \), we find the single step procedure proposed by Hommel and Hoffmann (1988) to control the PFER. In the two later cases, the procedure is a single step procedure and there is no need to order the p-values.

Theorem 4.6. For independent test statistics, the procedure defined above strongly controls the sEV at level \( \frac{m_0 \gamma}{m} \).

Proof. The proof of this theorem is a straightforward consequence of the following lemma.

Lemma 4.7. (Generalization of the main lemma in Benjamini and Hochberg (1995))

For any \( 0 \leq m_0 \leq m \) independent p-values corresponding to the true null hypotheses, and for any values that the \( m_1 = m - m_0 \) p-values corresponding to the false null hypotheses can take, the procedure 4.5 satisfies

\[
E(sFDP|p_{m_0+1}, \ldots, p_m) = \frac{m_0 \gamma}{m}.
\]

In Benjamini and Yekutieli (2001), a more general and simpler proof of the FDR control is provided. This proof can be generalized to prove the control of the sEV under positive dependency.

5. Simulations

When comparing multiple comparison procedures, one must set a common measure of performance and a common measure for the safeguards against false positives. We performed a limited simulation study in order to show the interest in using scaling functions that lie in between the FWER and the FDR. For this purpose, we use the average power as our measure of performance and the expected number of false positives as a measure of the safeguard against false rejections. We restricted our investigation to the case where the null distribution of the test statistic is a standard normal and under the alternative the distribution is a shifted normal with mean \( \Delta > 0 \) and variance 1. The number of tests \( m \) is either 1,000 or 10,000 and the number of alternative hypotheses (or false null hypotheses) is \( m_1 = \pi m \). The control levels are \( \alpha = 0.5 \) and \( \gamma = 0.05 \). For the sTP tests including the FER, the parameter \( \alpha = 0.5 \) means that the median of the sFDP is controlled to be less than \( \gamma \).

The two scaling functions we consider are members of the families mentioned before, that is,

\[
s_1(i) = \begin{cases} 
  i, & \text{if } i < k; \\
  k, & \text{if } i \geq k
\end{cases}
\]
Fig. 1. Average power and expected number of false rejections for \( m = 1,000 \) and \( \Delta = 2 \). Bonferroni (horizontal line), LR05 or BH95 (continuous line), sTP or sEV with \( s_1 \) (point dashed) and with \( s_2 \) (dashed). The control levels are \( \gamma = 0.05 \) and \( \alpha = 0.5 \).

\[ s_2(i) = i^\beta. \]

The constants of the scaling functions were chosen as \( k = 1/\gamma \) for \( m = 1,000 \) and \( k = 2.5/\gamma \) for \( m = 10,000 \). Note that as long as \( E(V) < k\gamma \), this scaling function gives the same control as the FDR, but will not allow \( E(V) \) to grow beyond \( k\gamma \). For the second function we chose \( \beta = 0.8 \).

The general error rate procedures were compared to some existing standards, Bonferroni, BH95 and LR05.

Figures 1 through 4 show the simulated average power and expected false positives \( E(V) \). The power of the Bonferroni procedure with \( \alpha = 0.05 \) can be computed analytically and is equal to

\[ \text{Pow}_{\text{Bonf}} = 1 - \Phi(\Phi^{-1}(1 - \frac{0.05}{m}) - \Delta) \] (13)

The construction of the figures is always the same. There are four panels, each plotting a performance or safety measure as a function of \( \pi \), the fraction of false null hypotheses. The top row of the panels contains the scaled tail probability procedures and uses the LR05 procedure as a standard. These are step-down tests. The bottom row shows the scaled expected value procedures and uses the BH95 for comparison. These are step-up tests. The left-hand column shows the average power (the performance indicator) and the right-hand column depicts \( E(V) \) (the safety indicator).

As expected, the scaled tests are in between the extremes defined by Bonferroni and BH95 or by LR05 since the scaling functions are chosen to be in between the horizontal line that corresponds to the Bonferroni procedure and the line with slope \( \gamma/m \) that represents...
Fig. 2. Average power and expected number of false rejections for \( m = 1,000 \) and \( \Delta = 4 \). Bonferroni (horizontal line), LR05 or BH95 (continuous line), sTP or sEV with \( s_1 \) (point dashed) and with \( s_2 \) (dashed). The control levels are \( \gamma = 0.05 \) and \( \alpha = 0.5 \).

Fig. 3. Average power and expected number of false rejections for \( m = 10,000 \) and \( \Delta = 2 \). Bonferroni (horizontal line), LR05 or BH95 (continuous line), sTP or sEV with \( s_1 \) (point dashed) and with \( s_2 \) (dashed). The control levels are \( \gamma = 0.05 \) and \( \alpha = 0.5 \).
Fig. 4. Average power and expected number of false rejections for $m = 10,000$ and $\Delta = 4$. Bonferroni (horizontal line), LR05 or BH95 (continuous line), sTP or sEV with $s_1$ (point dashed) and with $s_2$ (dashed). The control levels are $\gamma = 0.05$ and $\alpha = 0.5$.

BH95 or LR05. The performance of the scaled tests are close to the BH95 or the LR05 when the effect to be detected is clear-cut ($\Delta = 4$). However, they are much safer to use than BH95 and LR05, in particular, using the cut-off function which affords a direct choice of a threshold that the number of false positives must not exceed. In the two cases, $m = 1,000$ or 10,000, where the effect to be detected is small ($\Delta = 2$), the performance of the scaled tests is weaker in particular when compared to BH95 and LR05, but these procedure pay a heavy price in terms of false rejections as $\pi$ grows. Finally, these simulations demonstrates the importance of including the new general error rates which bridges the gap between the PFER and the FWER, and the FDR and the FER, especially when the number of tests performed is large in which case, the number of false positives generated by the FDR and FER control procedures becomes intolerable.

6. Conclusion

We introduced a new quantity, the scaled false discovery proportion sFDP, and we defined two metrics of false positives, the sTP and the sEV, that bridge the gap between the FWER and the FER and between the PFER and the FDR. For particular choices of the scaling function, the two metrics generalize the existent error rates. The new metrics offer to the user a large range of control by varying the scaling function depending on the choice of the level of conservativeness. We proposed as well some procedures to control either the sTP or the sEV under different assumptions. Other existent procedures can be generalized in the same way as presented in this paper. Two families of scaling functions were proposed to the user. We provided simulations that contrast the proposed procedures to standard ones in terms of power and control of false positives. The simulations showed the importance...
of including the new general error rates especially, when the number of tests is very large, which is often encountered in modern applications.

7. Appendix

Proof. Theorem (4.2) The proof is based on the method of Lehmann and Romano (2005). The event \( \text{sFDP} > \gamma \) occurs only if for at least one random index \( i \), the quantity \( \text{sFDP} \) exceeds \( \gamma \). Among these indexes, denote the smallest one by \( j \). Then \( P \left[ \text{sFDP} > \gamma \right] \leq P \left[ \text{such} \, P \text{into} \, \left\lfloor \gamma s(1) \right\rfloor \leq j \right] \). Because of the definition of \( \gamma s \), the range of the possible values of \( \gamma s(j) \) is divided into \( \left\lfloor \gamma s(1) \right\rfloor \leq j < \left\lfloor \gamma s(1) \right\rfloor + 1, \left\lfloor \gamma s(1) \right\rfloor + 1 \leq j < \left\lfloor \gamma s(1) \right\rfloor + 2, \ldots, \left\lfloor \gamma s(m) \right\rfloor \leq j < \left\lfloor \gamma s(m) \right\rfloor + 1 \).

Because of the definition of \( j \), we must have \( p(j) \leq \alpha_j, \) \( H(j) \) is true and \( \left\lfloor \gamma s(j) \right\rfloor + 1 \leq m_0 \). Therefore, \( P \left[ \text{sFDP} > \gamma \right] \leq P \left[ \left\lfloor \gamma s(1) \right\rfloor \leq j < \left\lfloor \gamma s(1) \right\rfloor + 1 \cup \left\lfloor \gamma s(1) \right\rfloor + 1 \leq j < \left\lfloor \gamma s(1) \right\rfloor + 2 \right. \cup \left. \ldots \cup \left\lfloor M - 1 \right\rfloor \leq j < M \right] \), with \( M = \min \{ \gamma s(m), m_0 \} \).

Let \( k - 1 \leq \gamma s(j) < k \) for \( k \in \{ \left\lfloor \gamma s(1) \right\rfloor + 1, \ldots, M \} \). Then \( p(j) = q(k) \leq \alpha_j \) because \( \frac{k - 1}{M - j} \leq \gamma \) and \( \frac{k - 1}{M - j} > \gamma \). This implies that \( H(j) \) is the \( k \)th rejected true hypothesis, and \( k \leq j \leq m - (m_0 - k) \) which implies that \( m_0 \leq m + k - j \). Therefore, if \( k - 1 \leq \gamma s(j) < k \), the event \( \{ \text{sFDP} > \gamma \} \) at step \( j \) implies that \( q(j) \leq \frac{k\alpha}{m_0} \). So,

\[
P \left[ \text{sFDP} > \gamma \right] \leq \sum_{k=\left\lfloor \gamma s(1) \right\rfloor+1}^{m} P \left[ q(k) \leq \frac{k\alpha}{m_0}, k - 1 \leq \gamma s(j) < k \right]
\]

\[
\leq \sum_{k=\left\lfloor \gamma s(1) \right\rfloor+1}^{M} P \left[ \bigcup_{k=\left\lfloor \gamma s(1) \right\rfloor+1}^{M} \left\{ q(k) \leq \frac{k\alpha}{m_0} \right\}, k - 1 \leq \gamma s(j) < k \right]
\]

\[
\leq P \left[ \bigcup_{k=\left\lfloor \gamma s(1) \right\rfloor+1}^{M} \left\{ q(k) \leq \frac{k\alpha}{m_0} \right\} \right].
\]

Part (ii) follows trivially.

Proof. Lemma (4.7) Our approach is based on the method of Benjamini and Hochberg (1995). The proof of this claim is by induction on \( m \). Note that when \( m_0 = 0 \), sFDP is identically 0. In this case, the claim is true for any value of \( m \). So, we treat the case \( m_0 \geq 1 \).

The case \( m = 1 \). Two cases.
1. If \( R = 0 \) then sFDP = 0.
2. If \( R = 1 \) then \( V = 1 \). This leads to

\[
s\text{FDP} = \begin{cases} 1/s(1) \text{ with probability } s(1) \cdot \gamma, \\ 0 \text{ with probability } 1 - s(1) \cdot \gamma. \end{cases}
\]

It follows that,

\[
s\text{EV} = E \left[ s\text{FDP} \right] = 1/s(1) \times s(1) \cdot \gamma + 0 \leq \frac{1}{1} \cdot \gamma = \frac{m_0}{m}. \gamma.
\]

The case \( m > 1 \). Suppose that the claim is true for any \( m' \leq m \). We have to show that the claim holds for \( m + 1 \).


Denote by \( q_{(1)}, \ldots, q_{(m_0)} \) the p-values that correspond to the true hypotheses and without loss of generality, denote by \( r_1, \ldots, r_{m_1} (m_1 = 1, \ldots, m + 1 - m_0) \) the ordered p-values that correspond to the false hypotheses. Define \( j_0 \) by

\[
j_0 = \max_{1 \leq j \leq m_1} \frac{s (m_0 + j)}{(m + 1)} \gamma.
\]

Here, \( j_0 \) is well defined because \( s \) is a non decreasing function. We set \( p' = \frac{s(m_0 + j_0)}{(m + 1)} \gamma \). \( q_{(m_0)} \) is either \( p' \) or \( \leq p' \). Then,

\[
E (\text{sFDP}|P_{m_0 + 1}, \ldots, P_m) = \int_0^{p'} E (\text{sFDP}|P_{m_0 + 1} = r_1, \ldots, P_m = r_m, q_{(m_0)} = p) f_{q_{(m_0)}}(p) \, dp
\]

\[
+ \int_{p'}^1 E (\text{sFDP}|P_{m_0 + 1} = r_1, \ldots, P_m = r_m, q_{(m_0)} = p) f_{q_{(m_0)}}(p) \, dp
\]

\[
= I + II
\]

with \( f_{q_{(m_0)}}(p) = m_0 p^{(m_0 - 1)} \).

In the first integral \( p \leq p' \), that is, \( m_0 + j_0 \) hypotheses, including the \( m_0 \) true hypotheses are rejected. Thus, \( \text{sFDP} = \frac{m_0}{s(m_0 + j_0)} \).

The first integral becomes

\[
I = \int_0^{p'} \frac{m_0}{s(m_0 + j_0)} \frac{m_0 p^{(m_0 - 1)}}{dp}
\]

\[
= \frac{m_0}{s(m_0 + j_0)} (p')^{m_0}.
\]

By the definition of \( j_0 \), we deduce that

\[
I = \frac{m_0}{s(m_0 + j_0)} \left( \frac{s(m_0 + j)}{(m + 1)} \right)^{m_0} = \frac{m_0}{(m + 1)} \gamma (p')^{m_0 - 1}.
\]

Now, for the second part. When both true and false hypotheses are considered together by their ordered p-values, the hypothesis \( H_i \) can be rejected only if there exists \( k, i \leq k \leq m_0 + j - 1 \), such that \( p(k) \leq \frac{s(k)}{m + 1} \gamma \), or equivalently

\[
\frac{p(k)}{p} \leq \frac{s(k)}{m_0 + j - 1} \frac{m_0 + j - 1}{p \cdot (m + 1)} \gamma.
\]

When conditioning on \( q_{(m_0)} = p \), each random variable \( q_i/p \), for \( i = 1, 2, \ldots, m_0 - 1 \), has a uniform \( U(0, 1) \) distribution. On the other hand, \( r_i/p \) for \( i = 1, \ldots, j \) are random variables situated between 0 and 1 (not necessarily of uniform distribution). Using the last inequality, to test \( m_0 + j - 1 \) hypotheses is equivalent to using the control procedure, with the constant \( \gamma' = \frac{m_0 + j - 1}{p \cdot (m + 1)} \gamma \). Applying the induction hypotheses, we have

\[
E (\text{sFDP}|P_{m_0 + 1} = r_1, \ldots, P_m = r_m, q_{(m_0)} = p) \leq \frac{m_0 - 1}{m_0 + j - 1} \frac{m_0 + j - 1}{p \cdot (m + 1)} \gamma = \frac{m_0 - 1}{p \cdot (m + 1)} \gamma.
\]

(14)
The bound in inequality (14) depends on $p$, but not on the segment $p_j < p < p_{j+1}$ for which it was evaluated, so

\[
II \leq \int_{p_j}^{1} \frac{m_0 - 1}{p \cdot (m + 1)} \gamma m_0 p^{(m_0 - 1)} dp \\
= \frac{m_0}{m + 1} \gamma \int_{p_j}^{1} (m_0 - 1) p^{(m_0 - 2)} dp \\
= \frac{m_0}{m + 1} \gamma \left\{1 - p^{(m_0 - 1)}\right\}.
\]

Finally,

\[
I + II \leq \frac{m_0}{(m + 1)} \gamma (p')^{m_0 - 1} + \frac{m_0}{m + 1} \gamma \left\{1 - p^{(m_0 - 1)}\right\} \\
= \frac{m_0}{(m + 1)} \gamma.
\]

References

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal* 52(6, SI), 708–721.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* 57(1), 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* 29(4), 1165–1188.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pub. del R Ist. Sup. di Sci. Eco. e Com. di Fir.* 8, 3–62. Bonferroni adjustment for multiple statistical tests using the same data.

Dudoit, S. and M. J. van der Laan (2008). *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. New York: Springer.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75(4), 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6(2), 65–70.

Hommel, G. and T. Hoffmann (1988). Controlled uncertainty. *Multiple Hypotheses Testing*, 154161.

Horn, M. and C. W. Dunnett (2004). Power and sample size comparisons of stepwise fwe and fdr controlling test procedures in the normal many-one case. *Lecture Notes-Monograph Series* 47, 48–64.

Lehmann, E. L. and J. P. Romano (2005). Generalizations of the familywise error rate. *Ann. Statist.* 33(3), 1138–1154.
Sarkar, S. K. (1998). Some probability inequalities for ordered mtp2 random variables: A proof of the simes conjecture. *The Annals of Statistics* 26(2), pp. 494–504.

Sarkar, S. K. (2007). Stepup procedures controlling generalized FWER and generalized FDR. *Ann. Statist.* 35(6), 2405–2420.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73(3), 751–754.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64(3), 479–498.

van der Laan, M. J., S. Dudoit, and K. S. Pollard (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.* 3, Art. 15, 27 pp. (electronic).

Vector, N. (1982). Exploratory data analysis and clinical research. *Methods of Information in Medicine* 21(2), 53–54.