# A Practitioner's Guide to Multiple Testing Error Rates

Jonathan Rosenblatt
Department of Statistics and Operations Research,
The Sackler Faculty of Exact Sciences,
Tel Aviv University
Israel

May 28, 2013

## 1 Introduction

It is quite common in modern research for a researcher to test many hypotheses. The statistical (frequentist) hypothesis testing framework does not scale with the number of hypotheses in the sense that naïvely performing many hypothesis tests will probably yield many false findings; "false" in the sense they will not be replicated. Indeed, classical statistical "significance" is evidence for the presence of a signal within the noise expected in a single test, not in a multitude. For protection from an uncontrolled number of erroneous findings, a researcher has to consider the type of errors, or non-replications, he wishes to avoid. The researcher can then select the adequate procedure for that particular error type and data structure, or alternatively estimate that error type for a particular set of candidate findings.

In practice, the selection of the proper error rate might cause the researcher some confusion. This point was made at the 2009 Multiple Comparisons conference in Tokyo [2, Section 4.4], demonstrated in the following question from the statistics Questions & Answers web site *Cross Validated* [1] :

[1] See `http://stats.stackexchange.com/questions/26588/multiple-fdr-corrected-experiments-using-the-same-data`. Accessed on Apr 20, 2013

*I am testing many (500,000) genetic variants, and the tests are FDR corrected and give me a q-value. Normally I would just call everything with* $q < .05$ *significant. But in this case I am testing those same genetic variants in two other related experiments (not using exactly the same individuals, but the samples may overlap). What to do? Would changing the significance threshold for* $q$ *to* $.05/3 = .0167$ *be an option?*

This particular example is further discussed in Section 4.2.

To offer guidance, we review possible error types for multiple testing (sec 2) and demonstrate them with some practical examples (sec 3) which clarify the formalism of sec 2. Finally, in appendix A, we include some notes on the software implementations of the methods discussed.

A multiplicity control procedure (e.g. Bonferroni, Benjamini-Hochberg, ...) is a data manipulation process– an algorithm– that guarantees that a preselected error rate is no larger than a preselected value. A typical procedures will actually offer guarantees vis-à-vis several error measures simultaneously. The emphasis of this manuscript is however on the error rates, and not on the multiplicity control procedures themselves.

For the purpose of selecting the appropriate procedure consult your favorite software's documentation (see our appendix A). Alternatively, Farcomeni [14] or more recently Goeman and Solari [18], can serve as references. As the focus of this paper is the error measures, p-value adjustment, simultaneous confidence intervals, and error estimation will not be discussed. The reader is reffered again to [14] or [18] as possible references.

# 2 Measures of Error

## 2.1 Family Wise Error Rates

Consider the testing of several null hypotheses against their respective research (alternative) hypotheses. The Family Wise Error Rate (FWER) is the frequency of experiments in which a false rejection of some null hypothesis will occur; put differently, the probability of a false finding.

As is customary in single hypothesis testing, a FWER level of $\alpha = 0.05$ is often used and sometimes even required, as in drug registering experiments.

Table 1 introduces the nomenclature which has become standard in the multiple comparisons community and will be referenced throughout this ar-

ticle. Following this notation, the FWER is defined as

$$Prob(V \geq 1)$$

where $Prob(.)$ denotes the relative frequency over repeated experiments.

| | Claimed nonsignificant | Claimed Significant | Total |
|---|---|---|---|
| Null | $U$ | $V$ | $m_0$ |
| Nonnull | $T$ | $S$ | $m_1$ |
| Total | $m - R$ | $R$ | $m$ |

Table 1: Classification of types of decisions made

### 2.1.1 Weak and Strong Control the Family Wise Error Rate

The probability of any particular inference procedure of making a false finding depends on the existence of true effects. FWER control in the "weak" sense refers to procedures which guarantee low FWER when there are no true effects at all, i.e., when all null hypotheses are correct.

Detecting the existence of *any* phenomena ($m_1 > 1$) is a simpler task that actually identifying these phenomena. Lack of weak FWER control means that we have no error guarantees even regarding this simple task. As mentioned in Section 1, a multiplicity control procedure might offer guarantees with respect to several error measure simultaneously. Weak FWER control should, and typically is, a minimal requirement.

FWER control in the "strong" sense is the complementing concept, referring to procedures which guarantee FWER control even in the presence of some true effects, i.e., when not all null hypotheses are correct. As their names imply, "strong" control is the stricter criterion, which entails "weak" control.

## 2.2 False Discovery Rates

The False Discovery Rate (FDR), first introduced by Benjamini and Hochberg [5], is the ratio between false discoveries and total discoveries, averaged over replicated experiments. Denoting the (unknown) False Discovery Proportion in a particular experiment using a particular data set as $FDP = V/R$, and

setting the convention that no discoveries signify no errors ($R = 0 \Rightarrow FDP = 0$), the FDR can be now defined as:

$$E\left(FDP\right)$$

where $E(.)$ denotes the average over all possible experimental results.

**Remark 2.1.** The FDR error rate has become synonymous with the Benjamini-Hochberg procedure presented in [5] . This is plain wrong and confusing. Benjamini-Hochberg is a procedure that does indeed offer FDR control in particular setups, but it is only one of many.

**Remark 2.2.** In the context of FDR, there need not be an $\alpha = 0.05$ convention. Researchers are free to choose the level of error they see adequate for their particular research. Obviously, an error level of $\alpha = 0.5$ might be hard to defend from critique. Having said that, since FDR control does offer FWER control in the weak sense, the $\alpha = 0.05$ convention might be justifiable.

## 2.3 Other Measures of Error

FWER and FDR are the most commonly used, but by no means the only measures of error. Since many error measures are merely an average over replications of the experiment, many other error measures can be considered by replacing the error function to be averaged. Denoting by $E(C)$ the average, over all possible experimental outcomes, of some error $C$. We now see that FWER and FDR simply set $C = I_{\{V \geq 1\}}$ and $C = FDP$ respectively. Some other measures of error are:

- Per Family Error Rate (PFER): Where $C = V$.
  This measure is the simple (expected) number of erroneous discoveries.
- Per Comparison Error Rate (PCER): Where $C = V/m$.
- $k$-FWER [28]: Where $C(k) = I_{\{V \geq k\}}$.
  This measure is the relative frequency of the making of no less than $k$ erroneous discoveries.
- False Discovery Exceedance (FDX)[17]: Where $C(\gamma) = I_{\{V/R \geq \gamma\}}$.
  This measure was motivated by the fact that FDR keeps the proportion

of false discoveries small, but only *on average*. In extreme scenarios, FDR does not exclude the possibility of making more than $FDP > \alpha$ mistakes in *almost all* experiments. FDX targets these scenarios explicitly, by allowing it to happen with a small probability, and is thus more conservative than FDR.

Other measures of error which are not simple averages over replications of the experiment include, but are not limited to:

- Positive FDR or pFDR [26] or $FDR_{-1}$ [1] : Defined as $E(V/R; R > 0)$. This measure is essentially the proportion of false findings (within all findings), but averaged, not on all possible experiments outcomes, but only on those which actually return findings. It was motivated by the observation that the FDR of a procedure might be very low merely because in many events it returns no findings, even if it makes many mistakes when it does indeed return findings (see [26] for a example). On the other hand, if all the null hypotheses are true, thus all findings are false, one would want an error measure to coincide with the probability of a false finding (weak FWER). pFDR does not enjoy this attribute.

- Marginal FDR or mFDR [27] or Fdr [12] or $FDR_{+1}$ [1]: Defined as $E(V)/E(R)$.
  While not very interesting for itself, this error measure gained popularity since it is mathematically tractable and approximates the FDR when many independent hypothesis are being tested.

We conclude by noting that FWER, FDR, pFDR and mFDR are (currently) by far the most popular. So much so that it is actually hard to find published applied research using any other.

## 2.4 Choosing Your Family

All the previous error measures are defined for a family of hypotheses that is known, and indirectly assumed this family is all the hypotheses being tested in an experiment. This need not be the case, and defining the family of relevance may be a non-obvious part of the researcher's work. The examples in Section 3 include some trivial scenarios, in the sense that the family of hypotheses is clear. The section also includes some examples where the family

is not trivial (see 4.3) and its choice will depend on the scientific statement in mind.

# 3 Examples

## 3.1 Tukey's Psychological Exams

In his 1953 unpublished paper: "The Problem of Multiple Comparisons" [4] and later, when lecturing at Princeton University [11], Prof. John Tukey would tell a motivating tale about a young psychologist. After administering 250 tests he finds that 11 were significant at the 0.05 level. A null hypothesis being that a given test does not differentiate between his groups of interest, a (significant) rejection of this null means a test does actually differentiate the groups. After the initial feeling of satisfaction he consults a senior researcher, only to discover his findings are rather poor, since one would expect 12.5 significant tests due to chance alone. Having only 11 significant results is actually disappointing.

With this new understanding, our psychologist now has to decide how he can protect himself from false findings. Say the tests consist of new candidate clinical diagnostics for condition X. Making an error means that a test will be used to diagnose X while it actually cannot distinguish between healthy and X. Since this is unacceptable for our psychologist, he will want an inference procedure that controls the FWER in the strong sense. This will also guarantee protection in the case that no test differentiates between healthy and X, i.e., weak FWER control, which can actually be a question of interest for itself.

Now consider a different scenario: The tests check for differences in personality attributes between genders. Making an error means that the psychologist might believe male and female differ in a way they actually do not. The researcher does not consider this a serious mistake, as long as many other true differences are discovered. In this setup, the researcher should control the FDR, FDX or pFDR. Allowing for some mistakes will allow the researcher to enjoy a sensitivity gain compared to FWER-controlling-procedures. The interpretation of the findings should be done in accord with the error measure employed.

## 3.2 ANOVA

In their 1999 paper, Williams, Jones, and Tukey, analyze the National Assessment of Educational Progress (NAEP) 1990 and 1992 data. This data consists of the average eighth grade mathematics proficiency scores for the 34 states that participated in both 1990 and 1992 NAEP Trial State Assessment (TSA). Comparisons are made between regions (Central, Northeast, Southeast and West), years (1990,1992), states nested within regions, and the Year $\times$ Region interaction. In this case a null hypothesis means there is no difference in the proficiency score between sub-groups, and its rejection meaning there is indeed such a difference.

We start by noting that this one study, provides four families of hypotheses. Indeed, a falsely discovered difference between regions, as an example, is of no concern when comparing years, states or regional changes.

We also note that in their paper, the authors actually compare between procedures controlling the FWER and the FDR. They do not offer a justification for the preference of any measure over the others, so we will offer one of our own. We will remark however, that their bottom line is unorthodox in the context of ANOVA:

> *Each of the three authors believes that the B-H procedure is the best available choice*

So why FWER? If, for example, the study is analyzed in the context of discrimination– having no policy implications but rather a possible stigmatizing effect– the researcher might wish to refrain from any falsely discovered differences between states, regions etc. If, on the other hand, intervention policies are the context, then power is a major concern. Missing a difference might mean policy makers are left unaware of the differences to be addressed. This context requires a less stringent error criterion than FWER, leading to the authors' stated preferences.

## 3.3 Functional Magnetic Resonance Imaging

Consider now the case of the neuroscientist, trying to locate the brain regions responsive to visual stimuli. He has scanned a dozen subjects or so in the Magnetic Resonance Imaging (MRI) machine and recorded the brain's activation[2] in response to the stimuli. To be precise, he measured the activation

[2] He actually measured the blood oxygenation level. Details can be found in [20].

level at *each* of several thousand brain locations, called Volumetric Picture Elements (voxels); their exact number depending on the resolution of the MRI scan . With the measured activation levels in hand, the researcher can compute their correlation to the stimulus given. If the voxel-wise measurement is correlated with the stimulus, the location is considered "active". We see that localizing activation actually consists of performing many local hypothesis tests: the null hypothesis of no correlation to the stimulus is tested at each voxel, and its rejection meaning a responsive location has been found.

Returning to multiplicity error rates; an error would mean the researcher declared a voxel as responsive when it actually is not. This does not seem like a terrible mistake to make, so the researcher should probably protect himself from large proportions of errors, and not from the making of one single error. FWER, k-FWER and Per Family Error Rate are thus excluded. Per Comparison Error Rate seems like a possible candidate, but it is very liberal. One can actually gain power by including many "junk" hypotheses, say, by including the air surrounding the head in the family. To see this, consider the case of infinitely many hypotheses tested. The proportion of errors will trivially be smaller than any $\alpha$ we pick.

Our researcher is thus left with FDR and FDX as candidates for measurement of error. In the case the researcher has no clear favorite from within these two measures, a possible consideration at this stage might be the availability, simplicity and power of controlling procedures. These considerations give preference, at the time of writing, to FDR over FDX.

**Remark 3.1.** For fairness it should be stated that "cheating" with FDR and FDR is possible, by augmenting the family with some obviously *false* null hypotheses [15]. It is our view that "cheating" the FDR or FDX does require more effort and malicious intent than analyzing a needlessly large brain volume. We thus do not qualify these two "cheats" as equally problematic.

### 3.3.1 Functional Magnetic Resonance Imaging– Cluster Level Inference

Return to the neuroscientist from sec 3.3. Recalling that the voxels are arbitrary volume units, defined by the technology of the MRI and not by entities of interest for inference, he decides that a more interesting entity is a mass of contiguous activations. He thus decides that he is interested in spatially contiguous regions with activation larger than "7" (in some scale).

These regions are known as "excursion regions", "exceedance sets", "blobs" and possibly other names. After scanning a subject, he realizes there are 30 contiguous regions which exceed 7. Conscious that some are due merely to chance variation, and knowing enough probability theory, he computes a p-value for the observed volume (exceeding 7), in each of the 30 regions. If he rather not make any mistakes, he can control the FWER of the regions. Namely, controlling the probability of declaring any inactive regions as active. This is indeed the approach implemented in several brain analysis software packages, particularly SPM (`http://www.fil.ion.ucl.ac.uk/spm/`).

Alternatively, if the researcher wishes to allow for some slack and accept false regions– as long as their proportion is not too high– he should use FDR or FDX control. Alas, the FDR defined in Section 2.2 assumes an a priori fixed and known number of hypotheses being tested ($m$). The number of excursion regions is data dependent, thus random and a priori unknown. Extensions of the FDR for the random-number-of-hypothesis case do exist. A rigorous exposition can be found in Siegmund et al. [23]. Note however, that error-controlling *procedures* for the random hypothesis case are not as abundant and studied as the fixed hypothesis case. The mathematical proofs would typically require some difficult to justify assumptions. Simulation performances however, do seem promising [9, 8].

### 3.4 Functional Magnetic Resonance Imaging– Clinical Scan

Return again to the neuroscientist from sec 3.3. This time his single patient is about to enter surgery for the removal of a brain tumor. The patient will be scanned in the fMRI in order to localize the speech regions, as the tumor is residing nearby and the surgeon needs to be extra-careful around these regions. In this clinical case there are different considerations than in basic scientific research. Type $II$ errors are arguably more important than type $I$ errors: underestimating the speech region might cost the patient his verbal skills; overestimating it, might cost him an extra surgery or a recurring tumor. None of the error measures presented until now is concerned with false negatives. Referring to the terminology in Table 1, our neuroscientist would probably be interested in something like $E(T/(m-R))$, which captures the sensitivity of the inference. This measure is the False Non Detection Rate (FNR) [16]. We have not presented this measure yet, as it is concerned with

the *non-detections*. We shall revisit it in the context of power in Section 6.

## 3.5 Genome Wide Association Studies

In a typical Genome Wide Association Study (GWAS) the geneticist will record the genetic information of many subjects (genotyping) with the aim of discovering associations between the genotype and the individuals' attributes (phenotype). Assuming a univariate phenotype, the researcher will perform some type of regression between the phenotype and each genetic attribute (titled single nucleotide polymorphism– SNP). With today's technology, the number of SNPs considered in a typical GWAS is hundreds of thousands. To declare an association, a researcher will try to reject the no association null hypothesis between *each* SNP and the phenotype, leading to the simultaneous testing of several hundreds of thousands of hypotheses. Since the researcher does not concern himself with the making of a single mistake, as long as other associations discovered are true, he should choose FDR control or one of its relatives discussed in Section 6. That said, it is also very common in GWAS, to use a p-value threshold of $10^{-7}$. This threshold is intended for FWER control, when searching over $500,000$ SNPs and using the Bonferonni procedure [7] for FWER control.

So FDR or FWER? It is left for the researcher to decide, and it ultimately depends on the implications of declaring false associations.

## 3.6 Cross Validated Example

In this example, the researcher is looking for associations between SNPs and three distinct[3] phenotypes. The error measure has already been selected. The family groupings are unclear. The options being (a) accounting for errors only within each experiment, leading to three families of hypotheses, or (b) global error accounting, leading to a single family.

Both approaches have their advantages and disadvantages. By keeping the experiments separate we gain power but the global error rate is no longer $\alpha$. Yekutieli [31] has approximated, for some cases, that under strategy (a) with $\alpha$ level FDR control within each experiment, then the *global* FDR should

[3] It is actually implicit whether these are distinct phenotypes or not. We have assumed they are distinct, because of the "subject overlap" comment.

actually be close to

$$FDR \approx \alpha \cdot \frac{\text{Total discoveries} + \text{No. of experiments}}{\text{Total discoveries} + 1} \tag{1}$$

Eq. 1 captures the intuition that the more experiments performed while controlling only for errors within the experiment, the global error rate might inflate.

The discussed problem includes three experiments, assumingly looking for three different phenomena. The fact we discuss *three* experiments, which were all conducted by the same researcher, is quite arbitrary. Why not control for the errors performed in the whole of science? Or at least in all genetic association studies. A proper discussion of this matter requires an unplanned detour into the philosophy and sociology of science, and is not part of this guide. We will conclude by remarking that combining errors over different phenomena is indeed desirable [19], yet rarely performed in practice.

## 3.7 Imaging Genetics

The field of imaging genetics aims at finding the genetic attributes associates with phenotypes derived from medical imaging. In a pioneering study, Stein et al. [24] set out to find the genetic variation associated with local brain volume, under the paradigm that different genes affect different brain regions. The data included the genotyping and imaging of $N \approx 700$ individuals. The genotype of each individual comprises information of $n_G \approx 400K$ SNPs. The imaging data encodes the relative volume of each subject at $n_B \approx 30K$ voxels. Testing for association between all $\{SNP\} \times \{voxel\}$ combinations, leads to $n_G \cdot n_B \approx 12 \times 10^9$ hypotheses. Should they all be considered one family of hypotheses? Or maybe each SNP (or voxel) is actually a separate family?

A researcher might want to infer which gene is associated with which location. A single family of hypotheses will include all $\{SNP\} \times \{voxel\}$ combinations. FWER control over this family is out of the question. FDR control means that the researcher is concerned with the proportion of false associations detected within all of the $\{SNP\} \times \{voxel\}$ associations found. This seems like a good criterion, except for the fact it requires correcting for $12 \times 10^9$ hypotheses. Our researcher starts considering alternative error criteria. A natural option might be SNP-wise testing, perhaps using the B-H procedure over all voxels within each SNP. Power is certainly gained as each family is corrected only for the $n_B$ voxels within it. What about false

findings? Sadly, this approach offers no error control. To see this, consider the case where there is only one voxel: this amounts to $n_G$ level $\alpha$ hypothesis tests, which is this initial multiplicity problem.

A more justifiable solution might harness the hierarchy of the problem; that is, by selecting associated SNPs and localizing the association only for these selected SNPs. Naturally, the multiplicity is alleviated since only selected SNPs will be passed for voxel-wise testing. However, if the same data is used for selecting the SNPs and then selecting the voxels, an $\alpha$ level FDR control within selected SNPs will still not guarantee an $\alpha$ level FDR control, across discovered $\{SNP\} \times \{voxel\}$ associations. This is actually a case of *selective inference* [2], also referred to by practitioners as "data snooping" or "double dipping".

Having given it more thought, our researcher decides she wants a method that has two properties: (a) controlling for the number of falsely discovered SNPs; and (b) controlling for the number of falsely discovered voxels associated with each discovered SNP.

To put if formally, Table 1 needs refinement. Define $R$ and $V$ to be the number of discovered SNPs and *falsely* discovered SNPS respectively. Define $R_g$ to be number of voxels declared associated with SNP $g$, and $V_g$ accordingly. The desired measure of error has two requirements:

$$E\left(\frac{V}{R}\right) \leq \alpha_1 \text{ and } E\left(\frac{1}{R}\sum_g \frac{V_g}{R_g}\right) \leq \alpha_2 \tag{2}$$

Is there a procedure that controls this type of error? While novel and under active research, there is presently one such procedure[4]. It has two stages. First, the omnibus-stage: testing for an associated SNP by aggregating over voxels within SNP and controlling for the number of SNPs tested. Second, a post-hoc stage: drilling into the selected SNPs searching for associated voxels. The novelty of the procedure is at the second stage, which controls for the number of voxels with a conservative error rate, which accounts for the previous SNP selection stage. The details can be found in [3].

[4] With proofs for the independent test-statistics case

# 4 Simultaneous versus Selective Inference

Up to this point, we have motivated the choice of error measure by a mere "error accounting". There is actually another perspective, which can make the choice of the error measure quite obvious once it has been recognized. As put by Cox [10]:

> *It might be better to talk about the problem of selected comparisons rather than about the problem of multiple comparisons.*

Cox's insight was that making statements on the truthfulness of a subset of selected hypotheses, and making statements about the simultaneous correctness of a subset of hypotheses are not the same thing. Think in terms of replicability: replicating a combination of phenomena is not the same as replicating each separately. Naturally, the simultaneous truthfulness of the selected hypotheses, entails the truthfulness of each and every one of them. Thus, simultaneous inference is the more ambitious task. In Cox's words [10]:

> *The fact that a probability can be calculated for the simultaneous correctness of a large number of statements does not usually make that probability relevant for the measurement of the uncertainty of one of the statements. If we are directly interested in a single statement about the vector parameter, the probability of simultaneous correctness would, however, be appropriate. The practical usefulness of the multiple comparison techniques then usually lies in giving a conservative bound for the effect of selection, rather than in giving an "exact" solution.*

Armed with the distinction between simultaneous and selective inference, we can relate error measures to inference types.

Simultaneity ambitions are only satisfied with FWER control. This is simply because allowing any errors to filtrate, ruins the truthfulness of the combination of claims. For the purpose of selective inference, one can consider FDR and its variants, which guarantee a small quantity of errors within the statements made, but not their simultaneity.

Demonstrating using the examples: The background in Tukey's psychological exams from Section 3.1 was too vague to determine which inference type is appropriate. Both can be advocated. The same goes for the NAEP

example from Section 3.2. Assuming the neuroscientist from the fMRI example in Section 3.3 cares of the truthfulness of each detected location, and not their particular combination, this is a case of selective inference. A similar consideration holds for the case of associated SNPs in the GWAS example in Section 4.1 and $\{SNP\} \times \{voxel\}$ associations in Section 4.3.

# 5 Power Considerations

The reader might have noticed that the different error measures in Section 2 care only of the number of discoveries and false discoveries. Our interest in detection sensitivity is naturally implicit in the procedures researchers employ. Otherwise, never rejecting any null hypothesis, will trivially control all of the error types in Section 2. Power can benefit from (a) knowledge of the proportion of signals in the noise $(1 - m_0/m)$ or (b) from an introduction the expected deviations from the null hypotheses.

To demonstrate (a), consider two researchers doing the same research. The first, which did some more reading on the topic, knows with complete certainty that there are 10 false null hypotheses (signals) in his 100 hypotheses tests. The second, being less through, has no access to this knowledge. Naturally, the first can exploit this information. As a trivial example, he will know that more than 10 rejections will certainly contain errors[5].

To demonstrate (b), consider a scenario where the researcher is certain that if an effect exists, it would be of magnitude $\pm 7$ (in some arbitrary scale, say z-values). Performing a one-sample $z$ or $t$ test, while ignoring this belief, will lead the researcher to reject all hypotheses with large (absolute) effects. Particularly, an effect of, say 20, will be considered very extreme, with infinitesimally small p-values. But, when considering the fact that effects are expected to be near 7, the researcher might actually prefer to reject effects near 7 *before* he rejects effects near 20. In statistical terminology, this is simply an underpowered test constructed for the wrong alternative hypothesis.

Specifying the expected deviation from the null for each hypothesis tested is no easy task. There exist however, several procedures which use the multitude of hypotheses tested in an attempt to empirically characterize the

[5] This does not mean that the first 10 are necessarily true. Only that more than 10 will certainly contain errors.

deviations from the null [6], and harness this information to gain power. Essentially all rely on estimators of the probability of a hypothesis being a true null given the value of some test statistic $z_i$. This is the posterior probability of the null, also named "local fdr" and denoted by $fdr(z_i)$. Details can be found in [12]. Also see Remark 6.1 on the existence and interpretation of this probability.

This magnitude– the probability of being null given the data– is not an error rate but rather a test statistic. It is however a rather intuitive test statistic. So much so that many authors set the rejection criterion to, say, $fdr(z_i) \leq 0.2$. This lends itself to the interpretation that results with frequencies smaller than 1 in 5, under the null assumption, are "dangerously prone to wasting investigators' resources" [12].

Storey [25] establishes a relation between $fdr(z_i)$ and the Marginal FDR from Section 2.3 so that a researcher opting for the $fdr(z_i) \leq 0.2$ criterion, can receive some sense of how many errors per discovery he will be doing on average. The relation is data dependent. In the problem analyzed by Efron [12], rejecting for $fdr(z_i) \leq 0.2$, is approximately equivalent to marginal FDR control of 0.1. This relation is even more appealing, since with a growing number of hypotheses being tested, the marginal FDR is a good approximation of the FDR.

Returning to power considerations, and using the notation presented in Table 1, we can specify many error measures which capture the idea of maximal power subject to the false detections being kept at a low level:

$$min\{FNR \quad \text{such that} \quad FDR \leq \alpha\} \tag{3}$$

$$min\{mFNR \quad \text{such that} \quad mFDR \leq \alpha\} \tag{4}$$

$$min\{T \quad \text{such that} \quad V \leq \alpha\} \tag{5}$$

The different procedures aimed at satisfying these error functions typically use the local fdr statistic ($fdr(z_i)$) as a test statistic, but differ in the heuristics used to compute this statistic [eg. 26, 12, 27], typically, assuming a large number of hypotheses being tested and independence between test statistics.

[6] Under mild assumptions regarding the form these deviations might take. Essentially assuming that deviations from the null are not uniformly dispersed but rather tend to clump together.

The search for procedures with desirable properties under errors measures such as eq 3 - eq 5, is ongoing. It is an active field of investigation with beautiful theory, developed either by the Multiple Comparisons Community or borrowing from statistical decision theory [see 27]. The analysis of the finite-sampling properties of suggested procedures with respect to some desirable error measures is not an easy task. For a complete treatment, and state of the art procedures, the reader is referred to [13].

**Remark 5.1.** Assigning a *posterior* probability to a hypothesis being null requires also assigning a *prior* probability to this event. The interpretation of this probability has raised some controversy [e.g. 21] as it can be seen as a statement of subjective beliefs, of sampling frequencies or merely as descriptive. Settling this interpretation issue is outside the scope of this manuscript.

# References


[1] Y. Benjamini. Discovering the false discovery rate. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 72(4):405–416, 2010.

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

[3] Y. Benjamini and M. Bogomolov. Adjusting for selection bias in testing multiple families of hypotheses. *Journal of the Royal Statistical Society. Series B (accepted)*, 2013.

[4] Y. Benjamini and H. Braun. John w. tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6):1576–1594, December 2002.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JOURNAL-ROYAL STATISTICAL SOCIETY SERIES B*, 57:289–289, 1995.

[6] F. Bretz, T. Hothorn, and P. Westfall. *Multiple Comparisons Using R*. Chapman and Hall/CRC, 1 edition, July 2010. ISBN 1584885742.

[7] W.S. Bush and J.H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12), December 2012.

[8] J. Chumbley, K. Worsley, G. Flandin, and K. Friston. Topological FDR for neuroimaging. *NeuroImage*, 49(4):3057–3064, 2010.

[9] J.R. Chumbley and K.J. Friston. False discovery rate revisited: FDR and topological inference using gaussian random fields. *NeuroImage*, 44 (1):62–70, January 2009.

[10] D. R. Cox. A remark on multiple comparison methods. *Technometrics*, 7(2):223–224, 1965.

[11] D. Donoho and J. S. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3):962–994, June 2004. WOS:000221981400005.

[12] B. Efron. Microarrays, empirical bayes and the two-groups model. *Statistical science*, 23(1):1–22, 2008.

[13] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, 1 edition, September 2010. ISBN 0521192498.

[14] A. Farcomeni. A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4):347–388, August 2008.

[15] Helmut Finner and M. Roters. On the false discovery rate and expected type i errors. *Biometrical Journal*, 43(8):985–1005, 2001.

[16] C.R. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal Of The Royal Statistical Society Series B*, 64(3):499–517, 2002.

[17] C.R. Genovese and L. Wasserman. Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101 (476):1408–1417, December 2006.

[18] Jelle J. Goeman and Aldo Solari. Tutorial in biostatistics: multiple hypothesis testing in genomics. *Statistics in Medicine*, 2013.

[19] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124 EP –, August 2005.

[20] N.A. Lazar. *The Statistical Analysis of Functional MRI Data*. Springer, 1 edition, July 2008. ISBN 0387781900.

[21] Carl N. Morris. Comment: Microarrays, empirical bayes and the two-groups model. *Statistical Science*, 23(1):34–40, February 2008. ArticleType: research-article / Full publication date: Feb., 2008 / Copyright © 2008 Institute of Mathematical Statistics.

[22] R Development Core Team. R: A language and environment for statistical computing. http://www.R-project.org, 2011.

[23] D. O. Siegmund, N. R. Zhang, and B. Yakir. False discovery rate for scanning statistics. *Biometrika*, 98(4):979 –985, December 2011.

[24] J.L. Stein, X. Hua, S. Lee, A.J. Ho, A.D. Leow, A.W. Toga, A.J. Saykin, L. Shen, T. Foroud, N. Pankratz, M.J Huentelman, D.W. Craig, J.D. Gerber, A.N. Allen, J.J. Corneveaux, B.M. DeChairo, S.G. Potkin, M.W. Weiner, and P.M. Thompson. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, 53(3):1160–1174, November 2010.

[25] J. D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *ANNALS OF STATISTICS*, 31(6):2013–2035, 2003.

[26] J.D. Storey. A direct approach to false discovery rates. *Journal Of The Royal Statistical Society Series B*, 64(3):479–498, 2002.

[27] W. Sun and T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, September 2007.

[28] 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*, 3:Article15, 2004.

[29] P.H. Westfall, R.R.D. Tobias, and R.D. Wolfinger. *Multiple Comparisons and Multiple Tests Using SAS, Second Edition*. SAS Institute, August 2011. ISBN 9781607648857.

[30] Valerie SL Williams, Lyle V. Jones, and John W. Tukey. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1):42–69, 1999.

[31] D. Yekutieli. Hierarchical false discovery Rate–Controlling methodology. *Journal of the American Statistical Association*, 103:309–316, March 2008.

# A On Your Computer

It does not suffice to choose an error measure in order to perform an analysis. An error-controlling procedure will also have to be chosen, and this is what you should look for in your favorite software. This might be a general purpose statistical suite, or a problem-specific application. In the latter case, we have little to suggest, as domain specific applications typically implement the procedures popularized in that field. In the brain imaging example, popular software include SPM, Brain Voyager, FSL, AFNII. All incorporate the multiplicity control procedure preferred by their authors (thus implicitly, the error measure). In the GWAS example, the same occurs in software such as Plink, PRESTO, PERMORY and others. General purpose statistical software are built for flexibility in the analysis, and thus incorporate more multiplicity control procedures.

In the R programming environment [22] the function *p.adjust* in the *stats* package will allow you to perform the most common procedures. The references in the function documentation are a good starting point for learning about these procedures. For FWER controlling procedures, in particular in the context of linear contrasts in regression models, the *multcomp* package is a good option. For FDR control (and variants) many packages have been written. A good listing of these can be found in Bretz et al. [6] or Korbinian Strimmer's web site: `http://strimmerlab.org/notes/fdr.html`. We also note that, to the best of our knowledge, the hierarchical testing scheme in Section 4.3 has not been implemented. Its implementation would require a new syntax to describe the hypotheses' hierarchy (families) and it is an open challenge.

In SAS, multiple testing procedures are incorporated within PROC MIXED and also in PROC MULTEST. The canonical reference is Westfall et al. [29].

In SPSS, multiplicity corrections are typically found as part of the *post-hoc* options of the analysis methods.

# B Glossary

Some of the terms in the multiple-comparisons literature, have appeared in several other disciplines under different names. To ease the transition, and for completeness, we present a glossary. In this glossary, we use as a reference the statistical nomenclature in Table 1. Also note that we use *rate* for the average of a *ratio* or a *proportion*. The literature is not consistent regarding this convention, so that the terms might be found in use for both purposes.

Table 2: Glossary

| Symbol | Names |
|---|---|
| $S$ | True Positives, Hits, True Discoveries |
| $U$ | True Negatives |
| $V$ | False Positives, Type $I$ Errors, False Discoveries, False Alarms |
| $T$ | False Negatives, Type $II$ Errors, False Non Discoveries, Misses |
| $V/R$ | False Discovery Proportion (FDP), False Discovery Ratio, False Detection Ratio/Proportion, False Alarm Ratio/Proportion, False Positive Ratio/Proportion, Fall-Out |
| $E(V/R)$ | False Discovery Rate (FDR), False Detection Rate, False Alarm Rate, False Positives Rate |
| $R/m$ | Accuracy |
| $E(T/(m-R))$ | False Non Discovery Rate (FNR) |
| $T/(m-R)$ | False Non Discovery Ratio |
| $S/R$ | Positive Predictive Value (PPR), Precision, Hit Ratio |
| $E(S/R)$ | Hit Rate |
| $U/(m-R)$ | Negative Predictive Value |
| $E(T)/m_1$ | Non Discovery Rate |
| $T/m_1$ | Non Discovery Ratio |
| $S/m_1$ | True Positive Ratio, Recall, Average Power |
| $U/m_1$ | Specificity, True Negative Ratio |
| $E(U)/m_1$ | True Negative Rate |
| $E(S/R)$ | True Positive Rate |