# A practitioner's Guide to Multiplicty Control

The Author

## 1 Introduction

It is quite common for modern research to test many hypotheses simultaneously. The frequentist hypothesis testing framework does not scale with the number of hypotheses in the sense that performing many $\alpha$ level tests will certainly yield many false findings. In order to scale, a researcher has to think of the type or errors he wishes to avoid and select the adequate method for that particular error type and data structure. A quick search of the tag *[multiple-comparisons]* in the statistics Questions & Answers web site Cross Validated, demonstrates the amount of confusion this task can actually cause. In an attempt to offer guidance at this important taks, we review some possible error types for simultaneous inference (sec 2) and demonstrate them with some examples (sec 3). We also include some notes on the software implementing these procedures in appendix A.

The emphasis of this manuscript is on the error rates, and not on the procedures themselves. We do however mention several procedures where appropriate. Simultanous confidence intervals and p-value adjustment and not discusses as they are procedure specific. I.e., it is the choice of a procedure that defines the p-value adjustment and the confidence intervals, and not the error rate itself.

## 2 Measures of Error

### 2.1 Family Wise Error Rates

Consider the testing of several null hypotheses against their respective research hypotheses. The Family Wise Error Rate (FWER) is the frequency of experiments in which a false rejection of a null hypothesis will occur. I.e.- a "false positive" finding will occur. Formally: Let $H_i^0$, for $i = 1, \dots, n$ be the family of null hypotheses, $T_i$ taking the value of 1 if the $i$'th null hypothesis is true and 0 otherwise, and $R_i$ taking the value 1 if the $i$'th hypothesis is rejected and 0 otherwise. The FWER is defined as $Prob\{\exists i : T_i = 1 \& R_i = 1\}$

## 2.2 False Discovery Rates

Consider the same setup as in the previous section. The False Detection Rate (FDR), first introduced by [cite BH 1995], is the average (over replicated experiments) ratio between false discoveries and total discoveries. Formally: Let $R = \sum_i R_i$ be the number of hypothesis rejected and $V$ be the number of *falsly* rejected hypotheses. I.e., $V = \sum_i R_i \cdot T_i$. Deciding that no rejections means no error[1], i.e. $V = 0 \Rightarrow V/R = 0$, then the FDR is defined as

$$E\left(\frac{V}{R}\right)$$

Remarks:

- In the context of FDR, the $\alpha \leq 0.05$ convention has not been fixed, and there is definite room for judgement as to the appropriate error rate for a specific problem.
- The FDR error rate has become synonymous with the Benjamini-Hochberg procedure. This is plain wrong as they should be distinguished.

## 2.3 Other Measures of Error

The measures of error above are the most commonly used, and for good reason, but by no means the only ones. Since an error measure is merely an expected loss over replications of the experiment, they can be expressed in the form $E(C)$ for some loss function $C$. In the cases of FWER and FDR $C = I_{\{V \geq 1\}}$ and $C = V/R$ respectively. Other measures of error include, but not limited to:

- Per Comparison Error Rate (PCER): Where $C = V/n$.

[1] From here onward, we will recall that $V = 0 \Rightarrow V/R = 0$ without mentioning it explicitly.

- Per Family Error Rate (PFER): Where $C = V$.
- False Discovery Exceedance (FDX)[cite Genovese and Wasserman 2006]: Where $C(\gamma) = I_{\{V/R \geq \gamma\}}$
- k-FWER [cite van der Laan 2004]: Where $C(k) = I_{\{V \geq k\}}$

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

- pFDR [cite Storey 2002] or $FDR_{-1}$ [cite Benjamini 2010] : Defined as $E(V/R; R > 0)$.
- Fdr [cite Efron 2008] or $FDR_{+1}$ [cite Benjamini 2010]: Defined as $E(V)/E(R)$.

### 2.4 Choosing Your Family

All the previous error measures assume that the family of hypotheses is known. As a researcher, defining the family is not an obvious task. The examples in section 3 included some trivial scenarios, in the sense the the family of hypotheses is obvious. The section also include some examples where the family is not trivial (sections 3.5 and 3.4) and it's choice will depend of the scientific statement the researcher is wishes to make.

## 3 Examples

### 3.1 Tukey's Psychological Exams

In his 1953 unpublished paper: "The Problem of Multiple Comparisons" [cite Bejnamini and Braun] and later, when lecturing at Princeton University [cite Donoho and Jin] a, John Tukey would tell the tale of a young psychologist. After administering 250 tests he finds that 11 were significant at the 0.05 level. After the initial feeling of satisfaction he consults a senior researcher, only to discover his findings are rather poor: One would expect 12.5 significant tests due to chance alone. Only 11 significant results is nothing to write home about.

With his new understanding, our psychologist now has to decide how should he 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 FWE. If Tukey's example, the tests are independent, so he could consider, say Sidak's procedure or Holm's procedure [what to cite? original papers not instructive enough. Hochberg's book? which page?].

Now consider a different scenario: where the tests check for differences in personality attributes between genders. Making an error means that the psychologist will believe male and female differ in a way they actually do not. The researcher does not consider this a serious mistake, as long as there are many other discovered differences. In this setup, the researcher should control the FDR– probably using the Benjamini-Hochberg method [cite BH 1995]. Allowing for some mistakes will allow the researcher to enjoy a sensitivity gain compared to FWER methods.

## 3.2 Functional Magnetic Resonance Imaging

Consider now the case of the neuroscientist, trying to locate the brain regions responding to visual stimuli. He will scan a dozen subjects or so in the Magnetic Resonance Imaging (MRI) machine and get the brain's activation[2] To be precise, he will be measuring the activation level at *each* of several thousand brain locations, called Volumetric Picture Elements (voxels). The exact number depending on the MRI's resolution. In the aim of finding "active" locations, the researcher will want to test the null hypothesis of "no response to the stimulus" at each voxel. A mistake would mean he declared a voxels 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, but not from the making of one single error. FDR is thus the error measure of choice.

## 3.3 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

[2] He actually measures the blood oxygenation level, but we leave the details for some other day.

discovering associations between the genotype and the individuals' attributes (phenotype). Assuming a univariate phenotype, the researcher will perform some sort of regression between the phenotype and each DNA location (SNP). With todays technology, the number of SNPs considered in a typical GWAS is in the hundreds of thousands. To declare an association, a researcher will try to reject the "no association" 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 mistake, as long as other associations discovered as true, he will choose FDR control. Probably using the Benjamini-Hochberg procedure.

## 3.4 Imaging Genetics

The field of imaging genetics, aims at finding the genetic attributes associates with phenotypes derived from medical imaging. In a pioneering study, [cite Stein] 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 about $N = 700$ individuals. The genotype of each individual comprises of information about close to $n_G = 400K$ DNA locations titled single nucleotide polymorphism (SNP). The imaging data encodes the relative volume at $n_B = 30K$ brain locations (voxels). The $n_G \cdot n_B = 1.2B$ hypotheses pose a computational challenge, but it is their logical structure that poses a conceptional challenge: what should be a considered as a family of hypotheses?

A researcher might want to infer which gene is associated with which location? She would then consider all SNP$X$voxel combinations and choose an appropriate multiplicity error measure; Probably FDR. She might want to detect which SNPs are associated with each voxel. In which case the association between *each* voxel and *all* SNPs is a family of hypotheses. She would thus control the error separately within each of the $n_V$ families. In the converse case, looking to detect which voxels are associated with each SNP, she will control the error within each of the $n_G$ hypotheses.

In the previous two scenarios, these is still a non-negligible chance that all findings are erroneous. This is because each family is considered as a separate problem. Having given it more thought- our researcher decides shes wants a method that controls for the number of falsely discovered SNPs and *simultaneously* for the number of falsely discovered voxels. More specifically- she wishes for FDR control over SNPs and for average (over discovered SNPs)

FDR control (over discovered voxels).

Formally The desired measure of error is

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

### 3.5 Electrode Selection in EEG

## 4 A Decision Theory Perspective

## 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 favourite software.

In the R programming environment [cite R Team] the function *p.adjust* in the *stats* package will allow 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 very good option. For procedures more sophisticated than Benjamini-Hochberg, many packages have been written. We will give references nor recommendations due to their fast rate of change.

R is special in that is decouples the statistical analysis and the multiplicity procedure via the *p.adjust* function. In SAS, multiple testing procedures are incorporated within PROC MIXED and also in PROC MULTEST. The canonical reference is [cite Westfall el al SAS Guide]. In SPSS, multiplicity corrections are typically found as part of the *post-hoc* options of the analysis methods.