FamilyWise Error Rate Control

Modern Inference

Aldo Solari

Acknowledgements

Much of the content was inspired by the following courses:

- Theory of Statistics by Prof. Emmanuel Cands
- Statistical methods for reproducibility by Prof. Aaditya Ramdas

and on a number of other sources.

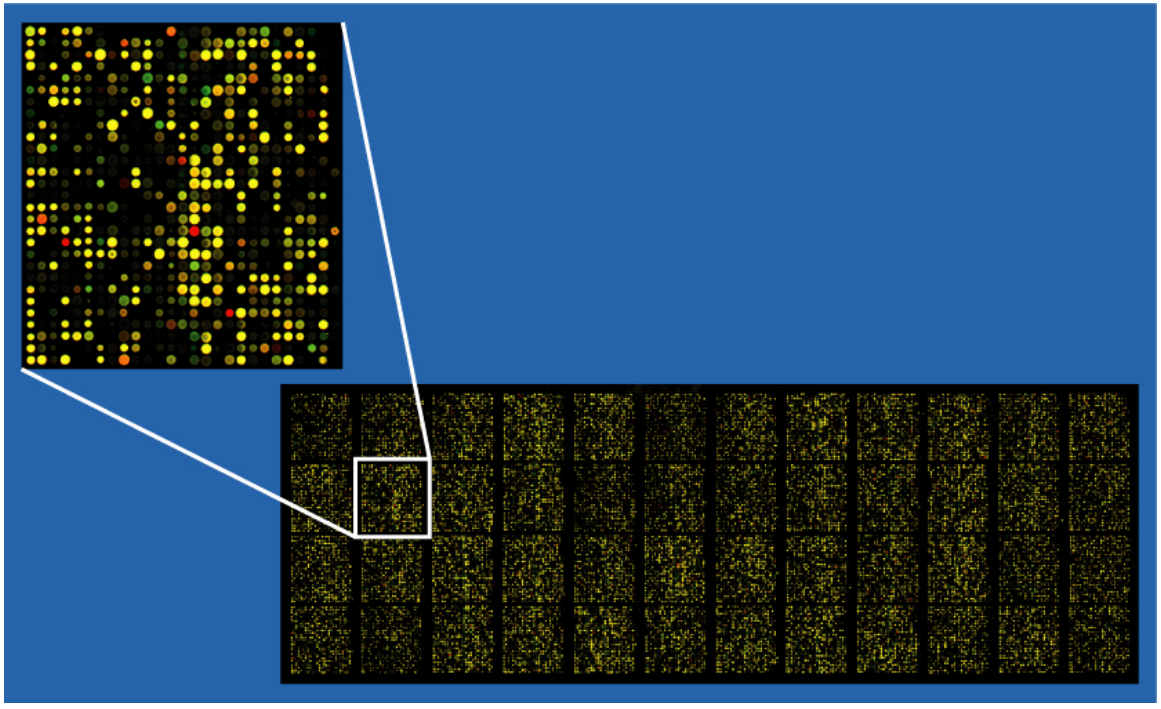
Outline

The three eras of statistics

Efron (2012)

1. **The age of huge census-level data sets** were brought to bear on simple but important questions:
   - Are there more male than female births?
   - Is the rate of insanity rising?
2. **The classical period** of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple
   - Is treatment A better than treatment B?
3. **The era of scientific mass production**, in which new technologies typified by the *microarray* allow a single team of scientists to produce *high-dimensional data*. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together

Microarrays



- Biomedical devices that enabled the assessment of individual activity for thousands of genes at once
- Need to carry out thousands of simultaneous hypothesis tests, done with the prospect of finding only a few interesting genes among a haystack of null cases

Prostate cancer data

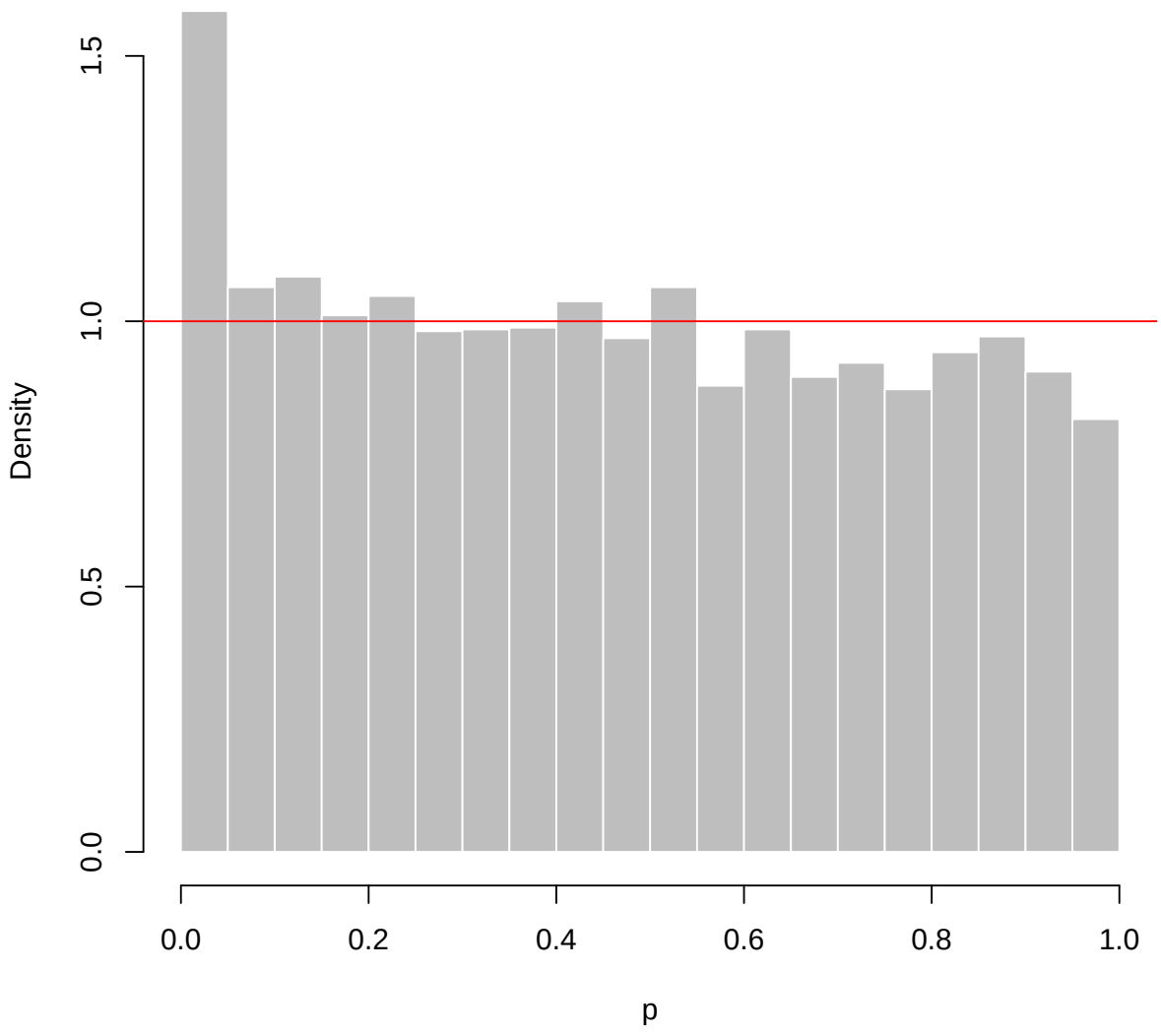
- The prostate cancer data came from a microarray study of $n = 102$ men, 52 prostate cancer patients and 50 controls
- Each man's gene expression levels were measured on $m = 6033$ genes, yielding a 102×6033 matrix

$$x_{ji} = \text{activity of } i\text{th gene for } j\text{th subject}$$

- The ith null hypothesis, denoted H_i, would state that the mean expression level of the ith gene is the same in both groups of patients

$$H_i : \mathbb{E}(X_i^{cancer}) = \mathbb{E}(X_i^{control})$$

- For each gene, a p-value p_i is computed


Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
p

The four eras of data

Leek (2016)

- **The era of not much data** Prior to about 1995, usually we could collect a few measurements at a time. The whole point of statistics was to try to optimally squeeze information out of a small number of samples - so you see methods like maximum likelihood and minimum variance unbiased estimators being developed
- **The era of lots of measurements on a few samples** This one hit hard in biology with the development of the microarray and the ability to measure thousands of genes simultaneously. This is the same statistical problem as in the previous era but with a lot more noise added. Here you see the development of methods for multiple testing and regularized regression to separate signals from piles of noise
- **The era of a few measurements on lots of samples** This era is overlapping to some extent with the previous one. Large scale collections of data from EMRs and Medicare are examples where you have a huge number of people (samples) but a relatively modest number of variables measured. Here there is a big focus on statistical methods for knowing how to model different parts of the data with hierarchical models and separating signals of varying strength with model calibration
- **The era of all the data on everything** This is an era that currently we as civilians don't get to participate in. But Facebook, Google, Amazon, the NSA and other organizations have thousands or millions of measurements on hundreds of millions of people. Other than just sheer computing I'm speculating that a lot of the problem is in segmentation (like in era 3) coupled with avoiding crazy overfitting (like in era 2)

Outline

Many tests

- In a single test, the probability of making a type I error is bounded by α, conventionally set at 0.05
- Problems arise, however, when researchers do not perform a single hypothesis test but many of them
- There are many ways of dealing with type I errors. We will focus on three types of multiple testing methods:
  1. those that control the *FamilyWise Error Rate* (FWER)
  2. those that control the *False Discovery Rate* (FDR)
  3. those that estimate the *False Discovery Proportion* (FDP) or make confidence intervals for it

Rejections

Suppose we have a collection $\mathcal{H} = \{H_1, \ldots, H_m\}$ of m null hypotheses:

- an unknown number m_0 of these hypotheses is true, whereas the other $m_1 = m - m_0$ is false. The proportion of true hypotheses is $\pi_0 = m_0/m$
- The collection of true hypotheses is $\mathcal{T} \subseteq \mathcal{H}$ and of false hypotheses is $\mathcal{F} = \mathcal{H} \setminus \mathcal{T}$
- The goal of a multiple testing procedure is to choose a collection $\mathcal{R} \subseteq \{1, \ldots, m\}$ of hypotheses to reject.
- If we have p-values $p_1, \ldots, p_m$ for $H_1, \ldots, H_m$, an natural choice is

$$\mathcal{R} = \{H_i : p_i \leq c\}$$

rejecting all hypotheses with a p-value below a critical value c

Errors

Ideally, the set of rejected hypotheses $\mathcal{R}$ should coincide with the set $\mathcal{F}$ of false hypotheses as much as possible. However, two types of error can be made:

- false positives, or type I errors, are the rejected hypotheses that are not false, i.e. $\mathcal{R} \cap \mathcal{T}$
- false negatives or type II errors are the false hypotheses that we failed to reject, i.e. $\mathcal{F} \setminus \mathcal{R}$

Rejected hypotheses are sometimes called *discoveries*, hence the terms *true discovery* and *false discovery* are sometimes used for correct and incorrect rejections

Type I errors

- Type I errors are traditionally considered more problematic than type II errors
- If a rejected hypothesis allows publication of a scientific finding, a type I error brings a false discovery, and the risk of publication of a potentially misleading scientific result
- Type II errors, on the other hand, mean missing out on a scientific result. Although unfortunate for the individual researcher, the latter is, in comparison, less harmful to scientific research as a whole
- Since each test has a probability of producing a type I error, performing a large number of hypothesis tests at α virtually guarantees the presence of type I errors among the findings

$$\mathbb{E}(\#\text{ type I errors}) = m_0 \times \alpha$$

2×2 **table**

We can summarize the numbers of errors in a contingency table:

	true	false	total
rejected	V	U	R
not rejected	$m_0 - V$	$m_1 - U$	$m - R$
total	m_0	m_1	m

We can observe m and $R = |\mathcal{R}|$, but all quantities in the first two columns of the table are unobservable

FDP

- Multiple testing methods try to reject as many hypotheses as possible while keeping some measure of type I errors in check
- This measure is usually either the number V of type I errors or the false discovery proportion (FDP) Q, defined as

$$Q = \frac{V}{\max(R, 1)} = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

the proportion of false rejections among all rejections, defined as 0 if no rejections are made

FWER and FDR control

Different types of multiple testing methods focus on different summaries of the distribution of V and Q. The most popular ones:

1. Familywise error rate (FWER):

$$\mathrm{FWER} = \mathrm{P}(V > 0) = \mathrm{P}(Q > 0)$$

   the probability that the rejected set contains any error

2. False discovery rate (FDR):

$$\mathrm{FDR} = \mathbb{E}(Q)$$

   the expected proportion of errors among the rejections

Either FWER or FDR is *controlled* at level α, which means that the set $\mathcal{R}$ is chosen in such a way that the corresponding aspect of the distribution of Q is guaranteed to be at most α, i.e.

$$\mathrm{FWER} \leq \alpha \quad \text{or} \quad \mathrm{FDR} \leq \alpha$$

FWER $\geq$ FDR

- The two error rates FDR and FWER are related. Because $0 \leq Q \leq 1$, we have $Q \leq \mathbb{1}\{Q > 0\}$ and

$$\mathrm{E}(Q) \leq \mathrm{P}(Q > 0)$$

  which means that FWER control implies FDR control
- Because FDR is smaller than FWER, we can generally expect FDR-based method to have more power than FWER-based ones, especially if there are many false hypotheses
- Conversely, if all hypotheses are true, FDR and FWER are identical; because $R = V$ in this case, Q is a Bernoulli variable, and

$$\mathrm{E}(Q) = \mathrm{P}(Q > 0)$$

- Both FDR and FWER are proper generalizations of the concept of type I error to multiple hypotheses: if $m = 1$, the two error rates are identical and equal to the type I error rate

Assumptions

- All methods we will consider start from a collection of test statistics $T_1, \ldots, T_m$, one for each hypothesis tested, with corresponding p-values

$$p_1, \ldots, p_m$$

  We call these p-values *raw* as they have not been corrected for multiple testing yet.

- Assumptions on the p-values often involve only the p-values of true hypotheses. We denote these *null* p-values by

$$q_1, \ldots, q_{m_0}$$

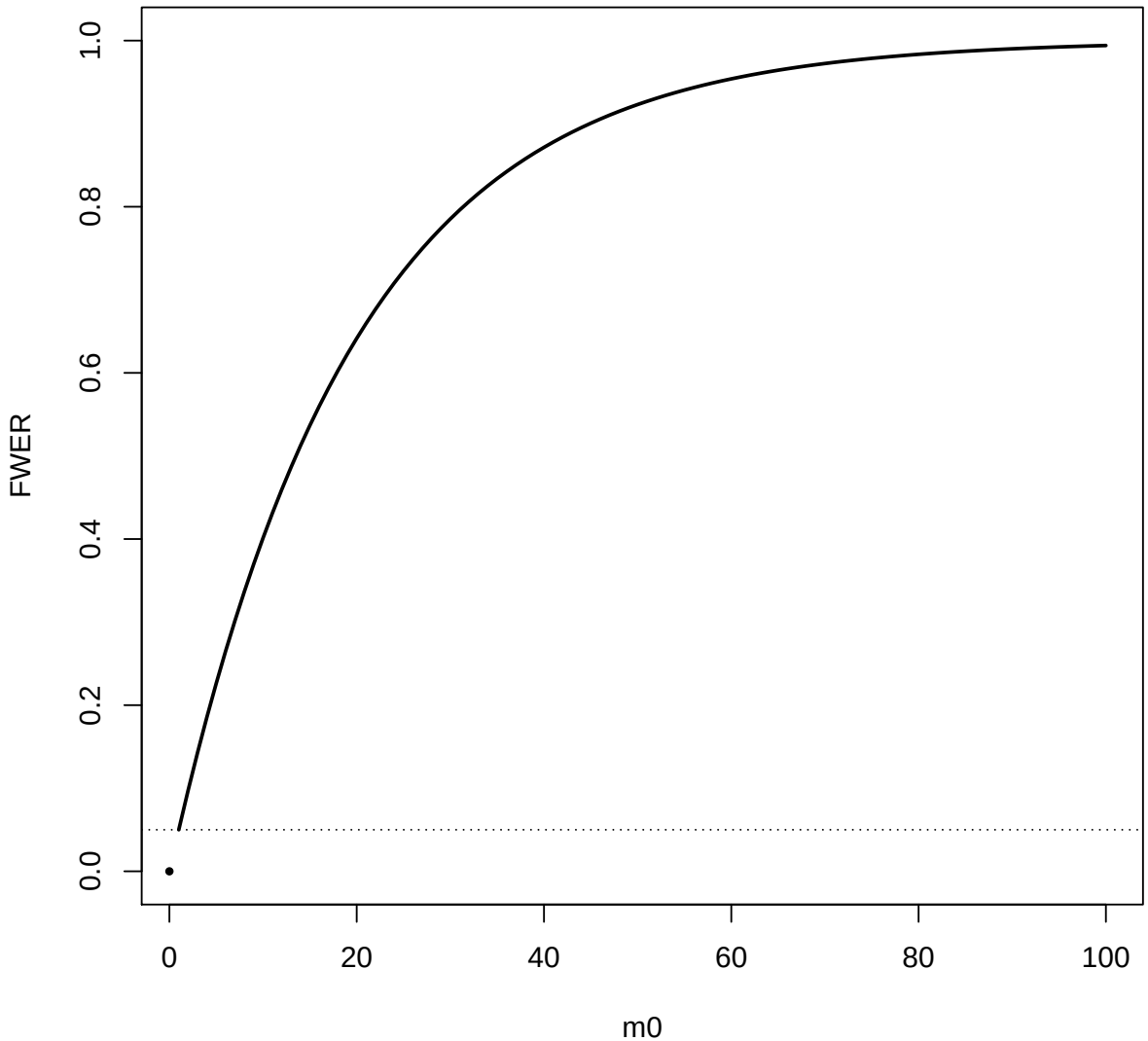
- Null p-values are assumed to be *valid* in the sense

$$\mathrm{P}(q_i \leq u) \leq u$$

  with equality when $q_i \sim U(0, 1)$

If $q_1, \ldots, q_{m_0}$ i.i.d. $U(0,1)$, then $\mathcal{R} = \{H_i : p_i \leq 0.05\}$ has

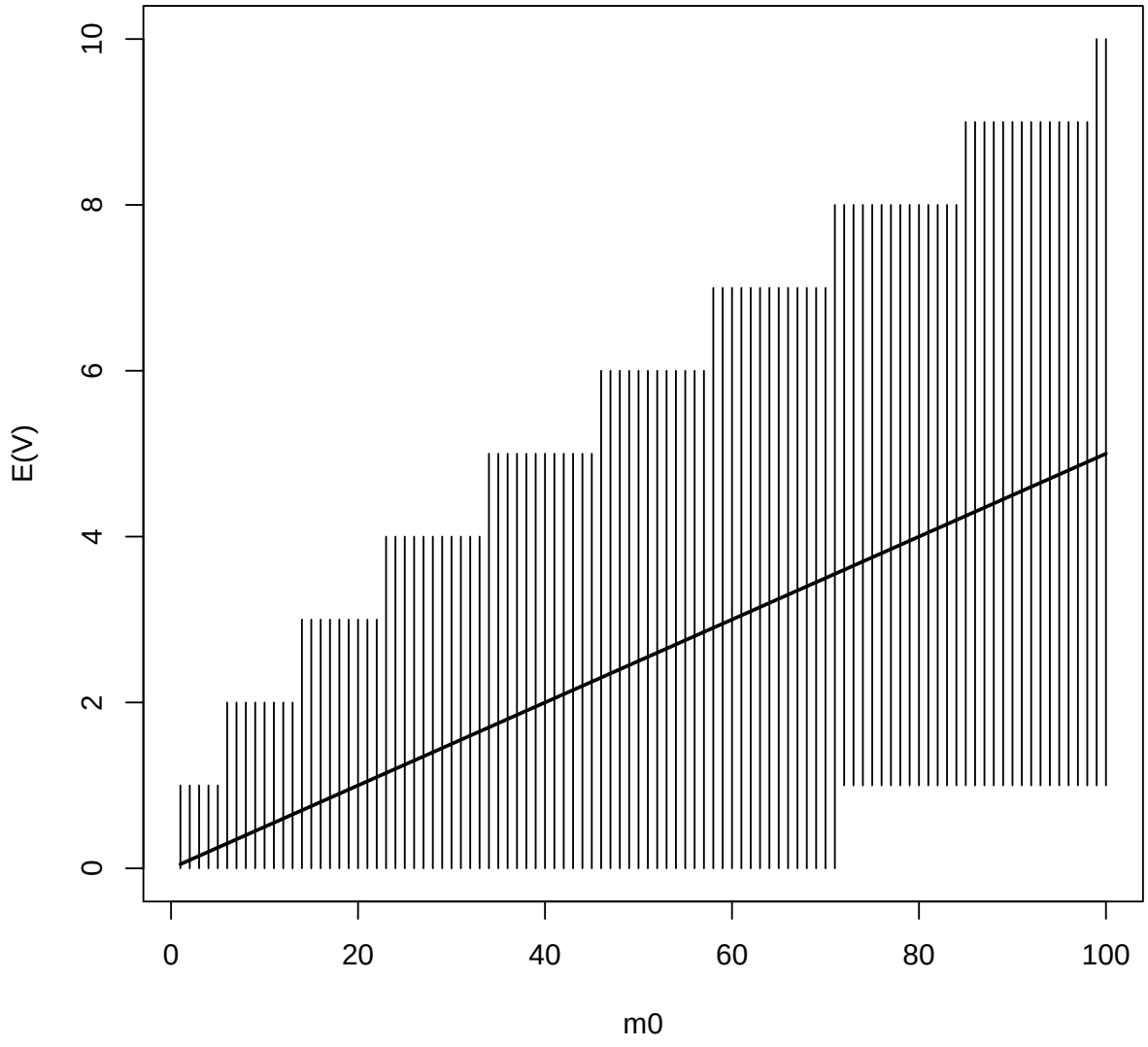
$$\text{FWER} = 1 - (1 - 0.05)^{m_0}$$



PFER

- The Per Family Error Rate (PFER) is the expected number of type I errors, i.e. $\mathrm{PFER} = \mathbb{E}(V)$. Take $\mathcal{R} = \{H_i : p_i \leq c\}$, and suppose that the null p-values $q_1, \ldots, q_{m_0}$ are marginally $U(0,1)$. Let $V = \sum_{i=1}^{m_0} \mathbb{1}\{q_i \leq c\}$
- $$\mathbb{E}(V) = \sum_{i=1}^{m_0} \mathbb{E}(\mathbb{1}\{q_i \leq c\}) = m_0 c$$
- $$\mathbb{V}\mathrm{ar}(V) = \sum_{i=1}^{m_0}\sum_{j=1}^{m_0} \mathbb{C}\mathrm{ov}(\mathbb{1}\{q_i \leq c\}\mathbb{1}\{q_j \leq c\}) = m_0 c(1-c) +$$
$$+2\sum_{i<j}\Big[\mathrm{P}(\mathbb{1}\{q_i \leq c, q_j \leq c\}) - \mathrm{P}(\mathbb{1}\{q_i \leq c\})\mathrm{P}(\mathbb{1}\{q_j \leq c\})\Big]$$
$$= m_0 c(1-c) + 2\sum_{i<j}\Big[\mathrm{P}(\mathbb{1}\{q_i \leq c, q_j \leq c\}) - c^2\Big]$$

  where the first term represents the independence structure and last term the *overdispersion*


E(V)
0
2
4
6
8
10
0
20
40
60
80
100
m0

PFER $\geq$ FWER

- By Markov's inequality

$$\mathrm{P}(V > 0) \leq \frac{\mathbb{E}(V)}{1}$$

  we obtain

$$\mathrm{FDR} \leq \mathrm{FWER} \leq \mathrm{PFER}$$

- Multiple testing methods are "identification" or "localization" procedures since they each generate m outputs instead of a binary reject/not reject the global null
- Identification implies detection, but not the other way around

Outline

Bonferroni method

The method of Bonferroni controls FWER at level α by rejecting hypotheses only if they have raw p-value smaller than α/m. Let $q_1, \ldots, q_{m_0}$ be valid p-values of the true null hypotheses

Theorem

Bonferroni method controls the FWER at level α:

$$\mathrm{FWER} \leq \pi_0 \alpha \leq \alpha$$

Proof.

$$\mathrm{P}\Big(\bigcup_{i=1}^{m_0} \big\{ q_i \leq \frac{\alpha}{m} \big\} \Big) \leq \sum_{i=1}^{m_0} \mathrm{P}\Big(q_i \leq \frac{\alpha}{m} \Big) \leq m_0 \frac{\alpha}{m} \leq \alpha.$$

□

Bonferroni conservativeness

The two inequalities indicate in which cases the Bonferroni method can be *conservative*, i.e. $\mathrm{FWER} < \alpha$

- The right-hand one shows that Bonferroni controls the FWER at level $\pi_0\alpha$, where $\pi_0 = m_0/m$. If there are many false null hypotheses, Bonferroni will be conservative
- The left-hand inequality is due to Boole's inequality, i.e. for any collection of events $E_1, \ldots, E_k$, we have $\mathrm{P}\big(\bigcup_{i=1}^{k} E_i\big) \leq \sum_{i=1}^{k} \mathrm{P}(E_i)$. This inequality is a strict one in all situations except the one in which all events $\{q_i \leq \alpha/m\}$ are disjoint. With independent p-values, the conservativeness is present but very minor. If $m_0 = m$ and $q_i \sim U(0,1)$, then

$$\mathrm{P}\Big(\bigcup_{i=1}^{m}\big\{q_i \leq \frac{\alpha}{m}\big\}\Big) = 1 - \Big(1 - \frac{\alpha}{m}\Big)^m \overset{m\to\infty}{\to} 1 - e^{-\alpha}$$

Sidak method

The method of Sidak rejects $\mathcal{R} = \{H_i \in \mathcal{H} : p_i \leq 1 - (1-\alpha)^{1/m}\}$ but it assumes independence of the null p-values.

Theorem

If the null p-values $q_1, \ldots, q_{m_0}$ are i.i.d. $U(0,1)$, Sidak method controls the FWER at level α.

Proof.

$\mathrm{P}\Big(\bigcup_{i=1}^{m_0}\{q_i \leq c\}\Big) = 1 - \prod_{i=1}^{m_0}\mathrm{P}\Big(q_i > c\Big) = 1 - (1-c)^{m_0}$ which equals α for $c = 1-(1-\alpha)^{1/m_0}$. Since we don't know m_0, we can use $1-(1-\alpha)^{1/m} \leq 1-(1-\alpha)^{1/m_0}$. □

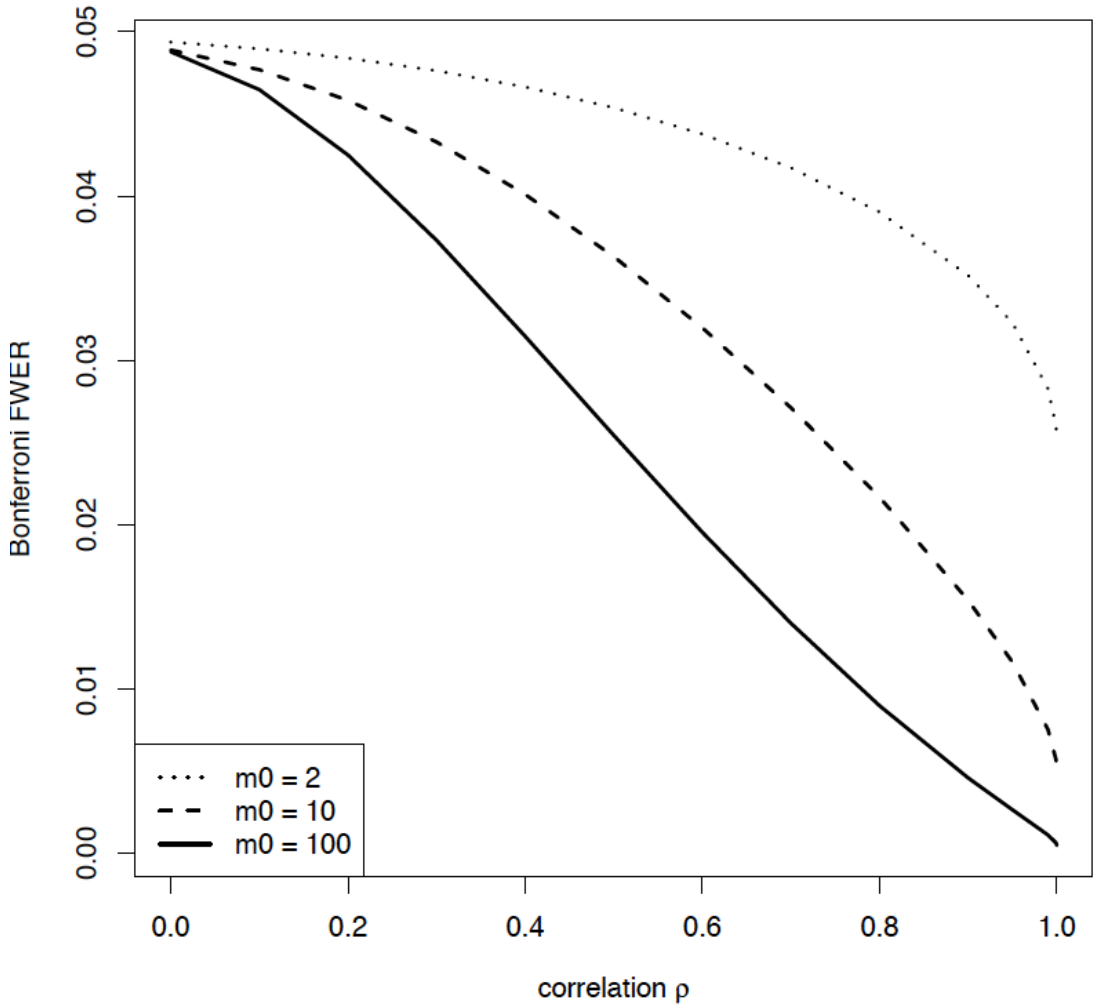
The ratio between the Bonferroni and Sidak critical values

$$\frac{\alpha/m}{1-(1-\alpha)^{1/m}} \overset{m\to\infty}{\to} \frac{-\log(1-\alpha)}{\alpha}$$

which evaluates to only 1.026 for $\alpha = 0.05$

Positively correlated p-values

Much more serious conservativeness can occur if p-values are positively correlated. Suppose that the correlation matrix is such that $\{\Sigma\}_{ij} = \rho$ for $i \neq j$



Adjusted p-values

- When testing a single hypothesis, we often do not only report whether a hypothesis was rejected, but also the corresponding p-value
- By definition, the p-value is the smallest chosen α-level of the test at which the hypothesis would have been rejected
- The direct analogue of this in the context of multiple testing is the *adjusted* p-value, defined as the smallest α level at which the multiple testing method would reject the hypothesis.
- For the Bonferroni procedure, this adjusted p-value is given by

$$\tilde{p}_i = \min(mp_i, 1)$$

where p_i is the raw p-value

Holm method

- Holm's method is a sequential variant of the Bonferroni method that always rejects at least as much as Bonferroni's method, and often a bit more, but still has valid FWER control under the same assumptions
- In the first step, all hypotheses with p-values at most α/h_0 are rejected, with $h_0 = m$ just like in the Bonferroni method. Suppose this leaves h_1 hypotheses unrejected. Then, in the next step, all hypotheses with p-values at most α/h_1 are rejected, which leaves h_2 hypotheses unrejected, which are subsequently tested at level α/h_2. This process is repeated until either all hypotheses are rejected, or until a step fails to result in any additional rejections

Holm algorithm

Step 0 Begin by ordering the p-values in ascending order

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$$

and let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ be the corresponding hypotheses

Step 1 : If $p_{(1)} \leq \alpha/m$ reject $H_{(1)}$ and go to Step 2. Stop otherwise

Step 2 : If $p_{(2)} \leq \alpha/(m-1)$ reject $H_{(2)}$ and go to Step 3. Stop otherwise

...

Step i : If $p_{(i)} \leq \alpha/(m-i+1)$ reject $H_{(i)}$ and go to Step $i+1$. Stop otherwise

...

Step m : If $p_{(m)} \leq \alpha$ reject $H_{(m)}$