False Discovery Rate
Part I : introduction et enjeux

E. Roquain

Laboratory LPMA, Université Pierre et Marie Curie (Paris 6), France

Point de Vue, 3rd February 2014
1 Introduction

2 False discovery rate control

3 FDR in other statistical issues
1 Introduction

2 False discovery rate control

3 FDR in other statistical issues
A “multiple testing joke” (http://xkcd.com)
A “multiple testing joke” (http://xkcd.com)
A “multiple testing joke” (http://xkcd.com)
A “multiple testing joke” (http://xkcd.com)

Multiplicity problem

\[ \Pr( \text{make at least one false discovery} ) \gg \Pr( \text{the } i\text{-th is a false discovery} ) \]

A correction is needed to assess significance!
Some other examples

Paradoxes due to large scale experiments

Probable facts appear significant
Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is \( R/(R + 1) \). The probability of a study finding a true relationship reflects the power \( 1 - \beta \) (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, \( \alpha \). Assuming that \( \epsilon \) relationships are being probed in the field, the expected values of the \( 2 \times 2 \) table are given in Table 1. After a research...

It can be proven that most claimed research findings are false.
Science-wise multiplicity issue - [Ioannidis (2005, PLoS Medicine)]

[Talk Benjamini Southampton (2013)] ; modeling:

Try many experiments

\[ \downarrow \]

1000 pure noise \quad 30 perfect signal

\[ \downarrow \]

publish results with a $p$-value \( \leq 0.05 \)

\[ \downarrow \]

\( \sim 50 \) false discoveries \quad 30 true discoveries

- Jager and Leek (2013). An estimate of the science-wise false discovery rate and application to the top medical literature \( \sim 14\% \)
- Ioannidis (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false
Science-wise multiplicity issue - [Ioannidis (2005, PLoS Medicine)]

[Talk Benjamini Southampton (2013)] ; modeling:

Try many experiments

\[\Downarrow\]

1000 pure noise \quad 30 perfect signal

\[\Downarrow\]

publish results with a \(p\)-value \(\leq 0.05\)

\[\Downarrow\]

\(\approx 50\) false discoveries \quad 30 true discoveries

- Jager and Leek (2013). An estimate of the science-wise false discovery rate and application to the top medical literature \(\approx 14\%\)

- Ioannidis (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false
Science-wise multiplicity issue - [Ioannidis (2005, PLoS Medicine)]

[Talk Benjamini Southampton (2013)] ; modeling:

Try many experiments

\[
\begin{align*}
&\downarrow \\
&1000 \text{ pure noise} & 30 \text{ perfect signal} \\
&\downarrow \\
&\text{publish results with a } p\text{-value } \leq 0.05 \\
&\downarrow \\
&\approx 50 \text{ false discoveries} & 30 \text{ true discoveries}
\end{align*}
\]

- Jager and Leek (2013). An estimate of the science-wise false discovery rate and application to the top medical literature \(\approx 14\%\)
- Ioannidis (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false
Science-wise multiplicity issue - [Ioannidis (2005, PLoS Medicine)]

[Talk Benjamini Southampton (2013)] ; modeling:

Try many experiments

\[\downarrow\]

1000 pure noise \hspace{1cm} 30 perfect signal

\[\downarrow\]

publish results with a $p$-value $\leq 0.05$

\[\downarrow\]

$\simeq 50$ false discoveries \hspace{1cm} 30 true discoveries

► Jager and Leek (2013). An estimate of the science-wise false discovery rate and application to the top medical literature $\simeq 14\%$

► Ioannidis (2014). Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false
Multiplicity in microarray [Hedenfalk et al. (2001)]

BRCA1 vs BRCA2

- expression level (activity)
- genes differentially activated?
- 1 test for each gene
- thousands of genes

- nb replications \(\ll\) dimension
- correlations
Other applications

- **Neuroimaging (FMRI)**
  activated regions?

- **Econometrics**
  winning strategies?

- **Astronomy**
  directions with stars?
Canonical setting

- $X_i = \text{avg group 2} - \text{avg group 1} \text{ (rescaled) for genes } i$
- Gaussian model:

$$
\begin{pmatrix}
X_1 \\
X_2 \\
\vdots \\
X_m
\end{pmatrix}
= \mu
\begin{pmatrix}
H_1 \\
H_2 \\
\vdots \\
H_m
\end{pmatrix}
+ \begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_m
\end{pmatrix},
$$

with $\mu > 0$, $H \in \{0, 1\}^m$ (fixed) and $\varepsilon \sim \mathcal{N}(0, \Gamma)$ ($\Gamma_{i,i} = 1$).

- $\Gamma = \text{dependence structure} = I_m$ for now

**Question:** for each $i$, $H_i = 0$ or $H_i = 1$?

Multiple testing: favors the “0” decision
Individual decision and errors

- Test statistic: \( X_i \)
- \( p \)-value: \( p_i = \Phi(X_i) \), with \( \Phi(z) = P(Z \geq z) \), \( Z \sim N(0, 1) \)

\( p_i \) such that

\[
\begin{align*}
\text{if } H_i = 0, & \quad p_i \sim U(0, 1) \\
\text{if } H_i = 1, & \quad p_i \sim \Phi(\Phi^{-1}(\cdot) - \mu)
\end{align*}
\]

- Choose \( \hat{H}_i = 1 \{ p_i \leq t \} \) for some threshold \( t \)

- Two errors:

<table>
<thead>
<tr>
<th>( H_i )</th>
<th>( \hat{H}_i = 0 )</th>
<th>( \hat{H}_i = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>true negative</td>
<td>false positive</td>
</tr>
<tr>
<td>1</td>
<td>false negative</td>
<td>true positive</td>
</tr>
</tbody>
</table>

- False positive more annoying
Picture 1. \( m = 100; \ m_0 = 50; \ \mu = 2; \ \Gamma = I_m \)
$m = 100; m_0 = 95; \mu = 3; \Gamma = I_m$
Picture 3. $m = 100; m_0 = 50; \mu = 0.01; \Gamma = I_m$
Picture 4. $m = 100; \ m_0 = 95; \ \mu = 0.01; \ \Gamma = I_m$
Doing like for 1 test? $t \equiv \alpha = 0.1$
Doing like for 1 test? \( t \equiv \alpha = 0.1 \)
Doing like for 1 test? $t \equiv \alpha = 0.1$
Union bound Bonferroni? \( t \equiv \frac{\alpha}{m} = 0.1/100 \)
Union bound Bonferroni? $t \equiv \alpha/m = 0.1/100$
Do something in between! $t_\ell = \frac{\alpha \ell}{m} = 0.1 \ell/100$
Do something in between! \( t_\ell = \alpha \ell / m = 0.1 \ell / 100 \)
Do something in between! $t_\ell = \alpha \ell / m = 0.1 \ell / 100$
Smart!
... and rigorous?
Introduction

False discovery rate control

FDR in other statistical issues
BH procedure

\[ \hat{k} = \max\{0 \leq i \leq m : p(i) \leq \alpha i / m\} \]

\[ \hat{t} = \max\{t \in [0, 1] : \hat{G}_m(t) \geq t / \alpha\} \]

\[ \hat{t} = \alpha \hat{k} / m \]

\[ \hat{H}_i = 1\{p_i \leq \hat{t}\} = 1\{X_i \geq \Phi^{-1}(\hat{t})\} \]
False discovery rate control

For a decision \( \hat{H}_i = 1 \{ p_i \leq \hat{t} \} (\forall i) \),

\[
\text{FDP}(\hat{t}) = \frac{\# \{ i : H_i = 0, \hat{H}_i = 1 \}}{\# \{ i : \hat{H}_i = 1 \}}
\]

\[
\text{FDR}(\hat{t}) = \mathbb{E}[\text{FDP}(\hat{t})]
\]

Theorem [Benjamini and Hochberg (1995)] [Benjamini and Yekutieli (2001)]

If \( \Gamma = I_m \) and \( \hat{t} \) threshold of BH procedure, \( \forall \mu, H \),

\[
\text{FDR}(\hat{t}) = (m_0/m)\alpha \leq \alpha
\]
Simulations. \( m = 50; \ m_0 = 25; \ \mu = 3; \ \Gamma = I_m \)
Simulations. \( m = 50; \ m_0 = 25; \ \mu = 3; \ \Gamma = I_m \)
Simulations. \( m = 50; \ m_0 = 25; \ \mu = 3; \ \Gamma = I_m \)
Simulations. $m = 50; m_0 = 25; \mu = 3; \Gamma = I_m$
Simulations. \( m = 50; \quad m_0 = 25; \quad \mu = 3; \quad \Gamma = I_m \)
Simulations. $m = 50; m_0 = 25; \mu = 3; \Gamma = I_m$
Simulations. $m = 50; m_0 = 25; \mu = 3; \Gamma = I_m$
Simulations. $m = 50; m_0 = 25; \mu = 3; \Gamma = I_m$
Simulations. $m = 50; m_0 = 25; \mu = 3; \Gamma = I_m$
Simulations. $m = 50; m_0 = 25; \mu = 3; \Gamma = I_m$
Benjamini and Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing

Impact

In many ways, Benjamini and Hochberg (1995) is a very successful paper. Its influence is clear from its 4967 citations (according to the Web of Science at the time of this session), which are still on the rise each year as can be seen in Fig. 1. Although 607 of these are in the area of statistics and probability, the majority of these publications are in the life sciences, from genetics to biochemistry, from oncology to plant sciences, reflecting in large part the use of FDR in microarray-related research. Importantly, citations in other high dimensional application areas, such as neural imaging, are on the rise also, showing its ability to be applied in many diverse types of application. The list of the 10 highest cited papers that cite Benjamini and Hochberg (1995), which is shown in Table 1, is particularly interesting, because it includes six statistical papers, suggesting that further theoretical and methodological developments of the method have had significant influence.

Fig. 1. Rapidly increasing number of citations of Benjamini and Hochberg (1995), suggesting that its influence has not yet reached its peak (note that the figure for 2009 is only partially shown)

Table 1. 10 most cited papers that cite Benjamini and Hochberg (1995)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Article citing Benjamini and Hochberg (1995)</th>
<th>Number of citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tusher et al. (2001)</td>
<td>3723</td>
</tr>
<tr>
<td>2</td>
<td>Storey and Tibshirani (2003)</td>
<td>1412</td>
</tr>
<tr>
<td>3</td>
<td>Weisberg et al. (2003)</td>
<td>1187</td>
</tr>
<tr>
<td>4</td>
<td>Genovese et al. (2002)</td>
<td>1020</td>
</tr>
<tr>
<td>5</td>
<td>Storey (2002)</td>
<td>726</td>
</tr>
<tr>
<td>6</td>
<td>Wilkinson (1999)</td>
<td>652</td>
</tr>
<tr>
<td>7</td>
<td>Benjamini and Yekutieli (2001)</td>
<td>584</td>
</tr>
<tr>
<td>8</td>
<td>Wacholder et al. (2004)</td>
<td>486</td>
</tr>
<tr>
<td>9</td>
<td>Patti et al. (2003)</td>
<td>479</td>
</tr>
<tr>
<td>10</td>
<td>Dudoit et al. (2002)</td>
<td>459</td>
</tr>
</tbody>
</table>

[Benjamini (2010, JRSSB)]

now > 20,000 citations on google scholar!
Introduction

False discovery rate control

FDR in other statistical issues
Why should FDR thresholding be adaptive to sparsity?
Why should FDR thresholding be adaptive to sparsity?
Why should FDR thresholding be adaptive to sparsity?
[Linnemann] - increasing signal strength

Bruit

Signal

Signal + Bruit
[Linnemann] - increasing signal strength
[Linnemann] - increasing signal strength
[Linnemann] - increasing signal strength
[Linnemann] - increasing signal strength
Adaptation to unknown sparsity

\( \hat{t} \) seems "adaptive" to the "quantity" of signal in the data

- Classification: where is the signal?
  [Bogdan et al. (2011)], [Neuvial and R. (2012)]

- Detection: is there some signal?
  [Ingster (2002)], [Donoho and Jin (2004)], etc

- Estimation: what is the value \( \hat{E}X \) of the signal?

  \[
  \hat{E}X = X_i \ 1\{|X_i| \geq \hat{t}\} \quad \text{(hard thresholding)}
  \]

  [Abramovich et al. (2006)], [Donoho and Jin (2006)]
$X_i \sim \pi_{0,m} \mathcal{N}(0, 1) + (1 - \pi_{0,m}) \mathcal{N}(\mu_m, 1), 1 \leq i \leq m, \text{ i.i.d.}$

but $\pi_{0,m} \rightarrow 1$ (sparse) and $\mu_m \rightarrow \infty$ (compensates sparsity).

- training set = null distribution known (one-class classification)
- Classification rule $\hat{h}_m : \mathbb{R} \rightarrow \{0, 1\}$;
- Risk

$$R_m(\hat{h}_m) = (1 - \pi_0)^{-1} \mathbb{E} \left( m^{-1} \sum_{i=1}^{m} 1\{\hat{h}_m(X_i) \neq H_i\} \right).$$

- Classification boundary in (sparsity, signal) space such that

Above the boundary, $\exists \hat{h}_m : R_m(\hat{h}_m) \rightarrow 0$ (perfect classification)

Under the boundary, $\forall \hat{h}_m, R_m(\hat{h}_m) \rightarrow 1$ (unclassifiable)
\[
\mu_m = \sqrt{2r \log m}
\]
\[
\pi_{0,m} = 1 - m^{-\beta}
\]
\[
\hat{h}_m^{BH}(x) = \mathbb{1}\{x \geq \Phi^{-1}(\hat{t})\}
\]

with \(\alpha_m \propto (\log m)^{-1/2}\)

Classification boundary attained by BH.

On the boundary: risk BH \(\sim\) Bayes risk.

[Bogdan et al. (2011)], [Neuvial and R. (2012)]
Detection: is there some signal?

Same model

\[ X_i \sim \pi_{0,m} N(0, 1) + (1 - \pi_{0,m}) N(\mu_m, 1), 1 \leq i \leq m, \text{ i.i.d.} \]

but \( \pi_{0,m} \to 1 \) (sparse) and \( \mu_m \to \infty \) (compensates sparsity).

- Test \( H_0 : "N(0, I_m)" \) against \( H_1 : "mixture". \)
  - Risk \( R_m(T) = P_{H_0}(T(X) = 1) + P_{H_1}(T(X) = 0) \)
  - Detection boundary in (sparsity, signal) space such that

    Above the boundary, \( \exists T : R_m(T) \to 0 \) (perfect detection)
    Under the boundary, \( \forall T, R_m(T) \to 1 \) (undetectable)
\( \mu_m = \sqrt{2r \log m} \)

\( \pi_{0,m} = 1 - m^{-\beta} \)

\( T^{BH} = 1 \{ \exists i : p(i) \leq \alpha_m i / m \} \)

with \( \alpha_m \propto (\log m)^{-1/2} \)

- Detection boundary attained by BH when \( \beta \in (3/4, 1) \)

- Better to use “higher criticism”

\[
\max_i \left\{ \sqrt{m} \frac{i/m - p(i)}{\sqrt{p(i)(1 - p(i))}} \right\}
\]
Regression with orthogonal design:

\[ X \sim \mathcal{N}(\beta, I_m) \]

[Bogdan et al. (2013)]: sorted \( \ell^1 \) penalized estimator (SLOPE)

\[
\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^m} \left\{ \frac{1}{2} \| X - \beta \|^2 + \sum_{k=1}^{m} \lambda_k \| \beta \|_k \right\}
\]

where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \); \( \| \beta \|_1 \geq \| \beta \|_2 \geq \cdots \geq \| \beta \|_m \)

Selection with \( \{ i : \hat{\beta}_i \neq 0 \} \):

- \( \lambda_k = \lambda = \Phi^{-1}(\alpha/(2m)) \approx \sqrt{2 \log m} \) \quad \text{Bonferroni}
- \( \lambda_k = \Phi^{-1}(\alpha k/(2m)) \approx \sqrt{2 \log(m/k)} \) \quad \text{BH !}
Regression with **orthogonal** design:

\[ X \sim \mathcal{N}(\beta, I_m) \]

[Bogdan et al. (2013)]: sorted \( \ell^1 \) penalized estimator (SLOPE)

\[
\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^m} \left\{ \frac{1}{2} ||X - \beta||^2 + \sum_{k=1}^{m} \lambda_k |\beta|^{(k)} \right\}
\]

where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m ; \ |\beta|^{(1)} \geq |\beta|^{(2)} \geq \cdots \geq |\beta|^{(m)} \)

Selection with \( \{ i : \hat{\beta}_i \neq 0 \} \):

- \( \lambda_k = \lambda = \Phi^{-1}(\alpha/(2m)) \simeq \sqrt{2 \log m} \) \quad \text{Bonferroni}
- \( \lambda_k = \Phi^{-1}(\alpha k/(2m)) \simeq \sqrt{2 \log(m/k)} \) \quad \text{BH !}
Outlook

Some conclusions for FDR

- Very simple
- Trade-off type I / power
- Adaptive to sparsity

Some issues

- Sensitive to null hypothesis
- Choosing $\alpha$
- Calibrating test statistics

Main challenge

What about dependence?
Outlook

Some conclusions for FDR

- Very simple
- Trade-off type I / power
- Adaptive to sparsity

Some issues

- Sensitive to null hypothesis
- Choosing $\alpha$
- Calibrating test statistics

Main challenge

What about dependence?
Some conclusions for FDR

- Very simple
- Trade-off type I / power
- Adaptive to sparsity

Some issues

- Sensitive to null hypothesis
- Choosing $\alpha$
- Calibrating test statistics

Main challenge

What about dependence?