FDR and Online FDR

Adel Javanmard and Andrea Montanari

USC and Stanford

December 11, 2015

Outline

1. Large-scale Hypothesis Testing
2. Controlling FDR
3. Controlling Online FDR
4. Conclusion

Large-scale Hypothesis Testing

Assume

- I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - Users are more likely to click on the first search result
  - Users are more likely to on top right ads
  - Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Assume

- I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - Users are more likely to click on the first search result
  - Users are more likely to on top right ads
  - Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Assume

- I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - Users are more likely to click on the first search result
  - Users are more likely to on top right ads
  - Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Assume

- I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - Users are more likely to click on the first search result
  - Users are more likely to on top right ads
  - Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Assume

- I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - Users are more likely to click on the first search result
  - Users are more likely to on top right ads
  - Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Assume

- I am the CTO of a big web company
- $\approx$ 1000 data scientists
- $\approx$ 1000 *'brilliant ideas'* per day
  - Users are more likely to click on the first search result
  - Users are more likely to on top right ads
  - Users are more engaged with page layout A
- How to avoid wasting company resources?

Compute 'significance level' from data!

Example

Idea: *Users click more on the first search result than on the second*

Null H_0: Users are equaly likely to click on first and second

Data:

- n events
- n_1 clicks on the *first* result
- $n_2 = n - n_1$ clicks on the *second* result

Idea

$$H_0 \quad \Rightarrow \quad z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1)$$

- If $z \gg 1$, then declare it significant

Example

Idea: *Users click more on the first search result than on the second*

Null H_0: Users are equaly likely to click on first and second

Data:

- n events
- n_1 clicks on the *first* result
- $n_2 = n - n_1$ clicks on the *second* result

Idea

$$H_0 \quad \Rightarrow \quad z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0, 1)$$

- If $z \gg 1$, then declare it significant

Example

Idea: *Users click more on the first search result than on the second*

Null H_0: Users are equaly likely to click on first and second

Data:

- n events
- n_1 clicks on the *first* result
- $n_2 = n - n_1$ clicks on the *second* result

Idea

$$H_0 \quad \Rightarrow \quad z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1)$$

- If $z \gg 1$, then declare it significant

Example

Idea: *Users click more on the first search result than on the second*

Null H_0: Users are equaly likely to click on first and second

Data:

- n events
- n_1 clicks on the *first* result
- $n_2 = n - n_1$ clicks on the *second* result

Idea

$$H_0 \quad \Rightarrow \quad z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1)$$

- If $z \gg 1$, then declare it significant

Formally

$$z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1)$$

p-value $(G \sim \mathrm{N}(0,1))$

$$p \equiv \mathbb{P}(G \geq z) = \int_z^\infty \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x$$

- Null: $p \sim \mathrm{Uniform}([0,1])$ (Definition)
- Small p: significant

Formally

$$z \equiv \frac{n_1 - n_2}{\sqrt{n}} \approx \mathrm{N}(0,1)$$

p-value $(G \sim \mathrm{N}(0,1))$

$$p \equiv \mathbb{P}(G \geq z) = \int_z^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x$$

- Null: $p \sim \mathrm{Uniform}([0,1])$ (Definition)
- Small p: significant

Company policy

Bring your idea up only if $p \leq \alpha$

[$\alpha = 0.05$, Fisher's rule of thumb]

Company policy

Bring your idea up only if $p \leq \alpha$

[$\alpha = 0.05$, Fisher's rule of thumb]

Problem

- $M \approx 1000$ hypotheses per day
- $M\alpha \approx 1000 \cdot 0.05 = 50$ pass the test
- Still too much waste

New company policy (Bonferroni):

Bring up your idea only if $p \le \alpha_M = \alpha/M$

Problem

- $M \approx 1000$ hypotheses per day
- $M\alpha \approx 1000 \cdot 0.05 = 50$ pass the test
- Still too much waste

New company policy (Bonferroni):

Bring up your idea only if $p \leq \alpha_M = \alpha/M$

Problem

- $M \approx 1000$ hypotheses per day
- $M\alpha \approx 1000 \cdot 0.05 = 50$ pass the test
- Still too much waste

New company policy (Bonferroni):

$$\text{Bring up your idea only if } p \le \alpha_M = \alpha/M$$

Problem with Bonferroni

Bring up your idea only if $p \leq \alpha_M = \alpha / M$

- More data scientists $\Rightarrow$ Less sensitive
- α false positives per day $\Rightarrow$ Does not scale with M

Problem with Bonferroni

Bring up your idea only if $p \leq \alpha_M = \alpha / M$

- More data scientists $\Rightarrow$ Less sensitive
- α false positives per day $\Rightarrow$ Does not scale with M

What do we want to achieve?

FDR (Benjamini, Hochberg, 1995)

- M hypotheses
- $\mathrm{D} \equiv$ Total number of discoveries (positives)
- $\mathrm{FD} \equiv$ Number of false discoveries

$$\mathrm{FDR} = \mathbb{E}\left\{\frac{\mathrm{FD}}{\max(\mathrm{D}, 1)}\right\}$$

Interpretation: $\mathrm{FDR} \le 0.1 \Rightarrow$ At most 10% of the discoveries is false.

FDR (Benjamini, Hochberg, 1995)

- M hypotheses
- $\mathrm{D} \equiv$ Total number of discoveries (positives)
- $\mathrm{FD} \equiv$ Number of false discoveries

$$\mathrm{FDR} = \mathbb{E}\Big\{\frac{\mathrm{FD}}{\max(\mathrm{D},1)}\Big\}$$

Interpretation: $\mathrm{FDR} \leq 0.1 \Rightarrow$ At most 10% of the discoveries is false.

Controlling FDR

Setting

Null hypotheses:

$$H_{0,1}, H_{0,2}, \dots, H_{0,M}$$

p-values:

$$p_1, p_2, \dots, p_M$$

Ground truth:

$$\theta_1, \theta_2, \dots, \theta_M [H_{0,i} : \theta_i = 0]$$

Test ouput ($\boldsymbol{p} = (p_i)_{1 \le i \le M}$:

$$T_1(\boldsymbol{p}), T_2(\boldsymbol{p}), \dots, T_M(\boldsymbol{p}) \in \{0, 1\}$$

$$\theta_i = 0 \Rightarrow p_i \sim \text{Uniform}([0, 1])$$

Setting

Null hypotheses:

$$H_{0,1}, H_{0,2}, \dots, H_{0,M}$$

p-values:

$$p_1, p_2, \dots, p_M$$

Ground truth:

$$\theta_1, \theta_2, \dots, \theta_M [H_{0,i} : \theta_i = 0]$$

Test ouput ($\boldsymbol{p} = (p_i)_{1 \le i \le M}$:

$$T_1(\boldsymbol{p}), T_2(\boldsymbol{p}), \dots, T_M(\boldsymbol{p}) \in \{0, 1\}$$

$$\theta_i = 0 \Rightarrow p_i \sim \text{Uniform}([0, 1])$$

Benjamini-Hochberg procedure

- Order the p-values

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(M)}$$

- Set threshold

$$I = \max\left\{ i \in [M] : \;\; p_{(i)} \le \frac{i\alpha}{M} \right\}$$

- Reject at level $p_{(I)}$:

$$T_\ell(p) = \begin{cases} 1 & \text{if } p_\ell \le p_{(I)}, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem (Benjamini, Hochberg, 1995)

If the p-values are independent, and BH is used, then

$$\mathrm{FDR} \le \alpha$$

Benjamini-Hochberg procedure

- Order the p-values

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$$

- Set threshold

$$I = \max\left\{i \in [M] : \quad p_{(i)} \leq \frac{i\alpha}{M}\right\}$$

- Reject at level $p_{(I)}$:

$$T_\ell(\boldsymbol{p}) = \begin{cases} 1 & \text{if } p_\ell \leq p_{(I)}, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem (Benjamini, Hochberg, 1995)

If the p-values are independent, and BH is used, then

$$\text{FDR} \leq \alpha$$

Interpretation

- M_0 true nulls, $M_1 = M - M_0$ true non-null
- Reject $H_{0,i}$ if $p_i \leq q$

$$\mathrm{FD} \approx M_0 q$$

$$\mathrm{D} = J(q) \equiv \max\{i : \; p_{(i)} < q\}$$

$$\mathrm{FDR} \approx \widehat{\mathrm{FDR}}(q) \equiv \frac{M_0 q}{J(q)} \leq \frac{Mq}{J(q)}$$

$$\widehat{\mathrm{FDR}}(p_{(I)}) \leq \frac{Mp_{(I)}}{I} \leq \alpha$$

Interpretation

- M_0 true nulls, $M_1 = M - M_0$ true non-null
- Reject $H_{0,i}$ if $p_i \leq q$

$$\mathrm{FD} \approx M_0 q$$
$$\mathrm{D} = J(q) \equiv \max\{i : \; p_{(i)} < q\}$$

$$\mathrm{FDR} \approx \widehat{\mathrm{FDR}}(q) \equiv \frac{M_0 q}{J(q)} \leq \frac{Mq}{J(q)}$$

$$\widehat{\mathrm{FDR}}(p_{(I)}) \leq \frac{M p_{(I)}}{I} \leq \alpha$$

Interpretation

- M_0 true nulls, $M_1 = M - M_0$ true non-null
- Reject $H_{0,i}$ if $p_i \le q$

$$\mathrm{FD} \approx M_0 q$$
$$\mathrm{D} = J(q) \equiv \max\{i : \; p_{(i)} < q\}$$

$$\mathrm{FDR} \approx \widehat{\mathrm{FDR}}(q) \equiv \frac{M_0 q}{J(q)} \le \frac{Mq}{J(q)}$$
$$\widehat{\mathrm{FDR}}(p_{(I)}) \le \frac{Mp_{(I)}}{I} \le \alpha$$

Controlling Online FDR

Back to our company

BH policy: Collect M p-values every day, and run BH

Problems:

- Centralized
- Controls end-of-day FDR
  Not end-of-year FDR

$\rightarrow$ Online FDR control

Back to our company

BH policy: Collect M p-values every day, and run BH

Problems:

- Centralized
- Controls end-of-day FDR
  Not end-of-year FDR

→ Online FDR control

Back to our company

BH policy: Collect M p-values every day, and run BH

Problems:

- Centralized
- Controls end-of-day FDR
  Not end-of-year FDR

$\rightarrow$ Online FDR control

Setting

Null hypotheses:

$$H_{0,1}, H_{0,2}, \dots, H_{0,M}$$

Sequence of p-values: one at each time

$$p_1, p_2, p_3, \dots$$

Ground truth:

$$\theta_1, \theta_2, \theta_3, \dots [H_{0,i} : \theta_i = 0]$$

Test ouput ($\boldsymbol{p}_1^t = (p_1, \dots, p_t)$):

$$T_1(\boldsymbol{p}_1^1), T_2(\boldsymbol{p}_1^2), T_3(\boldsymbol{p}_1^3), \dots \in \{0, 1\}$$

[Foster, Stine, 2007]

Setting

Null hypotheses:

$$H_{0,1}, H_{0,2}, \dots, H_{0,M}$$

Sequence of p-values: one at each time

$$p_1, p_2, p_3, \dots$$

Ground truth:

$$\theta_1, \theta_2, \theta_3, \dots [H_{0,i} : \theta_i = 0]$$

Test ouput ($p_1^t = (p_1, \dots, p_t)$):

$$T_1(p_1), T_2(p_2; T_1), T_3(p_3; T_1, T_2), \dots \in \{0, 1\}$$

[Foster, Stine, 2007]

What do we want to control?

- $\mathrm{FD}(n) \equiv$False discoveries up to time n
- $\mathrm{D}(n) \equiv$ Total number of discoveries up to time n

$$\mathrm{FDR}(n) \equiv \mathbb{E}\Big\{\frac{\mathrm{FD}(n)}{\max(\mathrm{D}(n), 1)}\Big\}$$

Want $\mathrm{FDR}(n) \leq \alpha$ for all n, θ

What do we want to control?

- $\mathrm{FD}(n) \equiv$False discoveries up to time n
- $\mathrm{D}(n) \equiv$ Total number of discoveries up to time n

$$\mathrm{FDR}(n) \equiv \mathbb{E}\Big\{ \frac{\mathrm{FD}(n)}{\max(\mathrm{D}(n), 1)} \Big\}$$

Want $\mathrm{FDR}(n) \leq \alpha$ for all n, θ

Trivial approach (Bonferroni)

- Choose $\beta_i \in [0,1]$, $\sum_{i=1}^{\infty} \beta_i \leq \alpha$
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \leq \beta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed

$$\mathrm{FDR}(n) \leq \mathbb{E}\{\mathrm{FD}(n)\} \leq \sum_{i:\theta_i=0} \mathbb{P}(p_i \leq \beta_i) = \sum_{i:\theta_i=0} \beta_i \leq \alpha$$

Very conservative!

Trivial approach (Bonferroni)

- Choose $\beta_i \in [0,1]$, $\sum_{i=1}^{\infty} \beta_i \leq \alpha$
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \leq \beta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed

$$\mathrm{FDR}(n) \leq \mathbb{E}\{\mathrm{FD}(n)\} \leq \sum_{i:\,\theta_i=0} \mathbb{P}(p_i \leq \beta_i) = \sum_{i:\,\theta_i=0} \beta_i \leq \alpha$$

Very conservative!

Trivial approach (Bonferroni)

- Choose $\beta_i \in [0,1]$, $\sum_{i=1}^{\infty} \beta_i \le \alpha$
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \le \beta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Indeed

$$\mathrm{FDR}(n) \le \mathbb{E}\{\mathrm{FD}(n)\} \le \sum_{i:\,\theta_i=0} \mathbb{P}(p_i \le \beta_i) = \sum_{i:\,\theta_i=0} \beta_i \le \alpha$$

Very conservative!

A simple rule

LORD (Levels based On Recent Discovery)

- Choose $\beta_i \in [0,1]$, $\sum_{i=1}^{\infty} \beta_i \leq \alpha$
- $\tau_i \equiv$ Time of the last discovery before i
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \leq \beta_{i-\tau_i}, \\ 0 & \text{otherwise.} \end{cases}$$

Each discovery resets everything.

A simple rule

LORD (Levels based On Recent Discovery)

- Choose $\beta_i \in [0,1]$, $\sum_{i=1}^{\infty} \beta_i \leq \alpha$
- $\tau_i \equiv$ Time of the last discovery before i
- Set

$$T_i = \begin{cases} 1 & \text{if } p_i \leq \beta_{i-\tau_i}, \\ 0 & \text{otherwise.} \end{cases}$$

Each discovery resets everything.

A theorem

Theorem (Javanmard, Montanari, 2015)

If the null p-values are indepenent, then LORD *satifies*

$$\sup_{\theta}\sup_{n} \mathrm{FDR}(n) \le \alpha\,.$$

Remarks

- Foster, Stine 2007:
  - Introduced model
  - Introduced *alpha investing rules*
  - Proved they control mFDR *(see next)*

- Last theorem applies to *generalized alpha investing*

- LORD uses very little information on the past!

Remarks

- Foster, Stine 2007:
    - Introduced model
    - Introduced *alpha investing rules*
    - Proved they control mFDR *(see next)*

- Last theorem applies to *generalized alpha investing*

- LORD uses very little information on the past!

Remarks

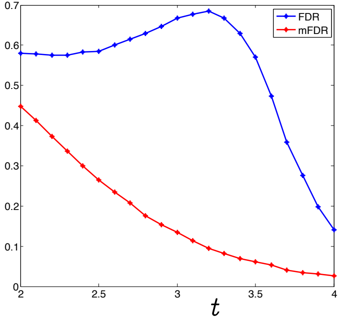
- Foster, Stine 2007:
  - Introduced model
  - Introduced *alpha investing rules*
  - Proved they control mFDR *(see next)*
- Last theorem applies to *generalized alpha investing*
- LORD uses very little information on the past!

FDRvs mFDR

$$\mathrm{mFDR}_\eta(n) = \frac{\mathbb{E}_\theta\{\mathrm{FD}(n)\}}{\mathbb{E}_\theta\{\mathrm{D}(n)\} + \eta}$$

$$\text{mFDR control} \not\Rightarrow \text{FDR control}$$

Example



Data

- $Z_1, \dots, Z_{n_0} \sim_{iid} \mathrm{N}(0,1)$, $(Z_{n_0+1}, \dots, Z_n) \sim \mathrm{N}(\theta_* \mathbf{1}, \rho \mathbf{1}\mathbf{1}^{\mathsf{T}} + \overline{\rho}\mathbf{I})$
- $n = 3000$, $n_0 = 2700$, $\theta_* = 2$, $\rho = 0.9$

Rule

$$T_i = \begin{cases} 1 & \text{if } |Z_i| \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

Statistical power?

Two-groups model

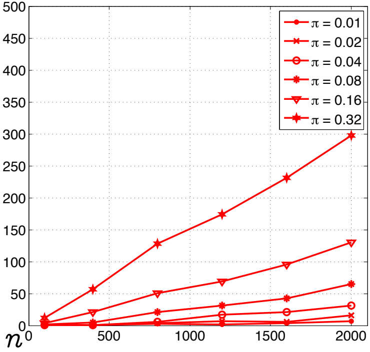
$$\theta_i \sim_{iid} \text{Bernoulli}(\pi)\,,$$

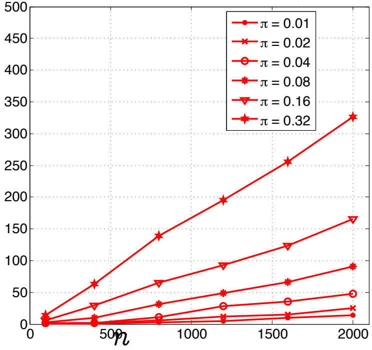
$$\mathbb{P}_{\theta_i}(p_i \le x) = \begin{cases} F(x) = x & \text{if } \theta_i = 0, \\ G(x) & \text{otherwise.} \end{cases}$$

'Discoveries should keep coming'

- A good rule should have $\mathrm{D}(n) = \Theta(n)$.

Two experiments





- **Left:** $\theta_i \sim_{iid} (1-\pi)\delta_0 + \pi \mathrm{N}(0,\sigma^2)$, $Z_i \sim \mathrm{N}(0,\theta_i)$
- **Right:** θ_i re-ordered, decreasing $|\theta_i|$

A theorem

Theorem (Javanmard, Montanari, 2015)

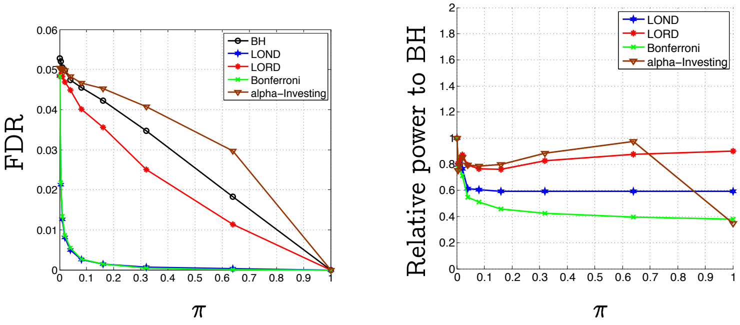
Assume the two-groups model, and use of LORD. *Then, almost surely*

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{D}(n) \geq \mathcal{A}(G,\beta),$$

$$\mathcal{A}(G,\beta) \equiv \left(\sum_{k=1}^{\infty} e^{-\sum_{\ell=1}^{k} G(\beta_\ell)}\right)^{-1}.$$

- $\mathcal{A}(G,\beta) > 0$ strictly if $G(\beta_\ell) > (1+\varepsilon)/\ell$ for all ℓ large enough.
- Sufficient $G(x) \approx G_0 x^{1+\delta}$ as $x \to 0$.

Comparison under the Gaussian two-groups model



$$\mathrm{TD}(n) = \text{ True discoveries}$$

$$\mathrm{RelativePower}(n) \equiv \mathbb{E}\left\{ \frac{\mathrm{TD}(n)}{\max(\mathrm{TD}_{\mathrm{BH}}(n), 1)} \right\}.$$

- $n = 1000$, $\sigma^2 = 2 \log n$

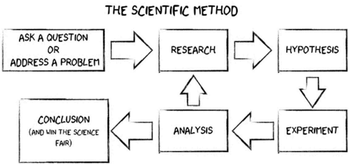
Conclusion

What if I am not CTO of a big-data company?

Take the “company” as a metaphor for science

RHETT ALLAIN SCIENCE 04.01.13 5:47 PM

WHAT'S WRONG WITH THE SCIENTIFIC METHOD?



More from The Economist | My Subscription

The Economist | World politics | Business & finance | Economics | Science & technology | Cultu

Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

Timekeeper | Like 28k | Tweet

PLOS | MEDICINE Browse

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

Conclusion

- FDR control is fundamental for reasoning about data

- Online FDR is likely more realistic

Thanks!

Conclusion

- FDR control is fundamental for reasoning about data

- Online FDR is likely more realistic

Thanks!

Conclusion

- FDR control is fundamental for reasoning about data
- Online FDR is likely more realistic

Thanks!