False Discovery Rate

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

1. **False Discovery Rate**

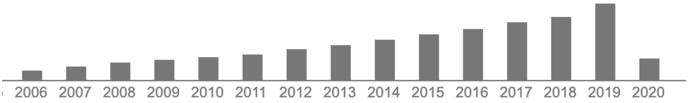
2. Benjamini & Hochberg

3. False Coverage Rate

False Discovery Rate

- Now we are testing millions of hypotheses at once, and making few false discoveries is not the end of the world
- The concept of False Discovery Rate (FDR) has changed thinking about multiple testing quite radically, showing that FWER control is not only way to do of multiple testing, and stimulating the field of multiple testing enormously
- FDR was introduced by Benjamini and Hochberg in 1995, and currently has 63K citations. It is one of the most-cited research of all time

- *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. Benjamini & Hochberg, JRSS-B (1995). Citations: 63K



- *Maximum likelihood from incomplete data via EM algorithm*. Dempster, Laird & Rubin, JRSS-B (1977). Citations: 60K
- *Nonparametric estimation from incomplete observations*. Kaplan & Meier, JASA (1958). Citations: 57K
- *Regression models and life-tables*. Cox, JRSS-B (1972). Citations: 51K

$m = 100$, $m_0 = 80$, $T_i \sim N(\mu_i, 1)$, $\mu_i = 0$ if H_i true, 2 otherwise

	1	2	3	4	5	6	7	8	9	10
R	20	17	23	16	20	16	15	17	20	17
V	4	5	6	5	5	3	3	5	7	4
$\mathbb{1}\{V > 0\}$	1	1	1	1	1	1	1	1	1	1
V/R	0.20	0.29	0.26	0.31	0.25	0.19	0.20	0.29	0.35	0.24

Reject H_i if $p_i \leq 0.05$, then FWER $= 0.983$ and FDR $= 0.232$

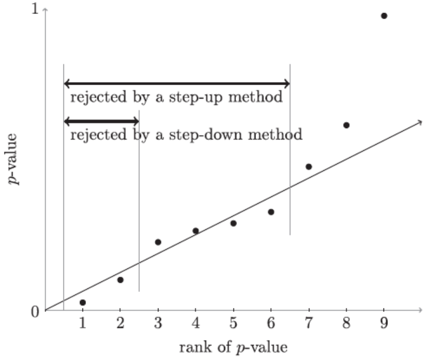
Outline

1. False Discovery Rate
2. **Benjamini & Hochberg**
3. False Coverage Rate

Benjamini & Hochberg

1. Sort the p-values
$$p_{(1)} \leq \cdots \leq p_{(m)}$$
2. If $p_{(i)} > \frac{i\alpha}{m}$ for all i, reject nothing, i.e. $\mathcal{R} = \emptyset$
3. Otherwise, let
$$i^* = \max\left\{ i \in \{1, \ldots, m\} : p_{(i)} \leq \frac{i\alpha}{m} \right\}$$
be the largest i for which $p_{(i)} \leq \frac{i\alpha}{m}$
4. Reject all $H_{(i)}$ with $i \leq i^*$, i.e. $\mathcal{R} = \left\{ H_i : p_i \leq \frac{i^*\alpha}{m} \right\}$

Step-up and step-down



- BH is a step-up method with threshold $\dfrac{i\alpha}{m}$
- Holm is a step-down method with threshold $\dfrac{\alpha}{m-i+1}$
- The ratio of thresholds is $\dfrac{i/m}{1/(m-i+1)} = i\Big(1 - \dfrac{i-1}{m}\Big)$

Theorem

For independent p-values $p_1, \ldots, p_m$ *and null p-values* $q_1, \ldots, q_{m_0} \overset{i.i.d.}{\sim} U(0,1)$, *the FDR of the Benjamini-Hochberg method is exactly* $\pi_0\alpha$, *i.e.*

$$\mathrm{FDR} = \pi_0\alpha \leq \alpha$$

Proof (Candes and Barber version)

- The conclusion is obvious when $m_0 = 0$: assume $m_0 \geq 1$
- Define $V_i = \mathbb{1}\{H_i \text{ rejected}\}$ for each $i \in T$ where $T = \{i : H_i \in \mathcal{T}\}$. We can express the FDP as

$$Q = \sum_{i \in T} \frac{V_i}{R \vee 1}$$

- We claim that

$$\mathbb{E}\Big(\frac{V_i}{R \vee 1}\Big) = \frac{\alpha}{m}, \quad i \in T$$

  based on which we have

$$\mathrm{FDR} = \mathbb{E}(Q) = \sum_{i \in T} \mathbb{E}\Big(\frac{V_i}{R \vee 1}\Big) = \sum_{i \in T} \frac{\alpha}{m} = \pi_0 \alpha$$

  What remains for the proof is to show that the claim is true

Proof - I

- When there are $R = k$ rejections, then H_i is rejected if and only if $p_i \leq (\alpha k)/m$, and therefore, we have

$$V_i = \mathbb{1}\{p_i \leq (\alpha k)/m\}$$

- Suppose $p_i \leq (\alpha k)/m$ (i.e. H_i is rejected). Let us take p_i and set its value to 0, and denote the new number of rejections by $R(p_i \downarrow 0)$. This new number of rejections is exactly R, because we have only reordering the first k p-values, all of which remain below the threshold $(\alpha k)/m$. On the other hand, if $p_i > (\alpha k)/m$, then we do not reject H_i, and so $V_i = 0$. Therefore we have

$$V_i \mathbb{1}\{R = k\} = V_i \mathbb{1}\{R(p_i \downarrow 0) = k\}$$

Proof - II

Combining the observations above and taking the expectation conditional on all p-values except for p_i, i.e. $\mathcal{F}_i = \{p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_m\}$, we have

$$\begin{aligned}
\mathbb{E}\Big(\frac{V_i}{R \vee 1}|\mathcal{F}_i\Big) &= \sum_{k=1}^{m} \frac{\mathbb{E}(\mathbb{1}\{p_i \leq (\alpha k)/m\}\mathbb{1}\{R(p_i \downarrow 0) = k\}|\mathcal{F}_i)}{k} \\
&= \sum_{i=1}^{k} \frac{\mathbb{1}\{R(p_i \downarrow 0) = k\}(\alpha k)/m}{k}
\end{aligned}$$

where the second equality holds because knowing $\mathcal{F}_i$ and $p_i = 0$ makes $\mathbb{1}\{R(p_i \downarrow 0)\}$ deterministic, and the fact that $p_i \sim U(0,1)$ and the p-values $p_1, \ldots, p_m$ are independent

Proof - III

- Next, we have

$$\mathbb{E}\Big(\frac{V_i}{R \vee 1}|\mathcal{F}_i\Big) = \frac{\alpha}{m}\sum_{k=1}^{m}\mathbb{1}\{R(p_i \downarrow 0) = k\} = \frac{\alpha}{m}$$

  after noticing that $\sum_{k=1}^{m}\mathbb{1}\{R(p_i \downarrow 0) = k\} = 1$
- Since we have set p_i to 0, we must make at least one rejection - we will always reject H_i. Therefore $R(p_i \downarrow 0) \geq 1$, and $R(p_i \downarrow 0)$ must take a value between 1 and m
- The tower property verifies that

$$\mathrm{FDR} = \sum_{i \in T}\mathbb{E}\Big(\frac{V_i}{R \vee 1}\Big) = \sum_{i \in T}\mathbb{E}\Big[\mathbb{E}\Big(\frac{V_i}{R \vee 1}|\mathcal{F}_i\Big)\Big] = \sum_{i \in T}\frac{\alpha}{m} = \pi_0\alpha$$

$\square$

Assumption of *positive regression dependence on a subset* (PDS). Formally, let $D \subset [0,1]^m$ be increasing set if $x \in D$ and $x \leq y \leq 1$ (in the coordinate-wise sense) together imply $y \in D$.

Definition

A set of p-values $(p_1, \ldots, p_m)$ is said to satisfy the PDS property if for any increasing set $D \subset [0,1]^m$ and each null index $i \in T$, the probability

$$\mathrm{P}((p_1, \ldots, p_m) \in D | p_i \leq t)$$

is non-decreasing in $t \in (0,1]$.

Examples of cases under which the PDS condition holds include one-sided test statistics that are jointly normally distributed, if all correlations between test statistics are positive.

Theorem

For p-values satisfying the PDS assumption, the Benjamini-Hochberg procedure controls the FDR at level α, i.e.

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

Adaptive Benjamini-Hochberg

- The Benjamini & Hochberg method, like Bonferroni, controls its error rate at level $\pi_0\alpha$, rather than at α. This suggests the possibility of an alternative, more powerful Benjamini & Hochberg procedure that uses critical values

$$\frac{i\alpha}{\hat{\pi}_0 m}$$

  rather than $(i\alpha)/m$ if a good estimate $\hat{\pi}_0$ of the proportion of true hypotheses π_0 would be available
- Such procedures are called *adaptive* procedures, and many have been proposed on the basis of various estimates of π_0
- A problem with the adaptive approach, however, is that estimates of π_0 can have high variance, especially if *p*-values are strongly correlated. Naive plug-in procedures, in which this variance is not taken into account, will therefore generally not have FDR control

π_0 estimator

$$\hat{m}_0(\lambda) = \frac{\sum_{i=1}^{m} \mathbb{1}\{p_i > \lambda\}}{1-\lambda}$$

If null p-values have marginal $U(0,1)$ distribution, a proportion $1-\lambda$ is expected to be above λ:

$$\mathbb{E}(\sum_{i=1}^{m} \mathbb{1}\{p_i > \lambda\}) \geq \mathbb{E}(\sum_{i \in T} \mathbb{1}\{q_i > \lambda\}) = m_0(1-\lambda)$$

thus $\mathbb{E}(\hat{m}_0) \geq m_0$.

$\hat{\pi}_0 = \dfrac{\hat{m}_0}{m}$ is a conservative estimator of π_0, i.e. $\mathbb{E}(\hat{\pi}_0) \geq \pi_0$

Storey method

1. Choose $\lambda \in (0,1)$. Estimate π_0 by

$$\hat{\pi}_0(\lambda) = \frac{\sum_{i=1}^{m} \mathbb{1}\{p_i > \lambda\} + 1}{(1-\lambda)m}$$

2. Perform Benjamini-Hochberg procedure at level

$$\frac{\alpha}{\hat{\pi}_0(\lambda)}$$

- The addition of 1 to the numerator makes sure that $1/\hat{\pi}_0$ is always well-defined (but it may happen $\hat{\pi}_0 > 1$)
- The value of λ is typically $1/2$, although $\lambda = \alpha$ has also been advocated
- Storey method controls FDR under independence of p-values but generally not under positive dependence

General dependence

1. It has been proved that the Benjamini-Hochberg procedure controls FDR at level α also under the PDS assumption
2. If the PDS assumption is not valid, an alternative is the procedure of Benjamini & Yekutieli, which is valid under general dependence

Example

Consider $m = m_0 = 2$. The two null p-values q_1 and q_2 are marginally $U(0,1)$, but the joint distribution of (q_1, q_2) is piecewise constant with density

- $1/(1-\alpha)$ in areas b
- $2/\alpha$ in area c
- $b(1 - b\alpha/2)$ in area a
- 0 in gray areas

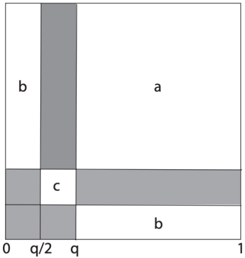


Figure 2: A piecewise constant joint distribution.

Benjamini & Yekutieli

$$\mathrm{FDR}[\mathrm{BH}(\alpha)] = \mathrm{P}(I) + \mathrm{P}(II) + \mathrm{P}(III)$$

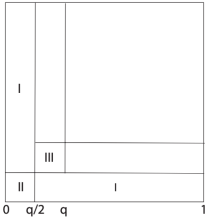


Figure 1: The BHq rejection region.

- In our example, $\mathrm{FDR}[\mathrm{BH}(\alpha)] = 3\alpha/2$
- In general, $\exists$ a worst-case joint distribution of p-values such that $\mathrm{FDR}[\mathrm{BH}(\alpha)] = \alpha H_m$ with $H_m = \sum_{j=1}^{m} \frac{1}{j}$ harmonic number
- Benjamini & Yekutieli method uses $\dfrac{i\alpha}{mH_m}$ to control FDR at α

Outline

Simultaneous confidence intervals

- Suppose that we have m parameters of interest $\theta_1, \ldots, \theta_m$ and we wish to construct a confidence interval for each of them such that *all* the intervals will contain their respective parameters with probability at least $1 - \alpha$, i.e.

$$\mathrm{P}(\bigcap_{i=1}^{m} \{\widetilde{\mathrm{CI}}_i \ni \theta_i\}) \geq 1 - \alpha$$

- For example, with $y_1, \ldots, y_m$ are independent with $y_i \sim N(\theta_i, 1)$, marginal $1 - \alpha$ confidence intervals

$$\mathrm{CI}_i = [y_i - z_{1-\alpha/2}, y_i + z_{1-\alpha/2}]$$

  are such that $\mathrm{P}(\mathrm{CI}_i \ni \theta_i) = 1 - \alpha$
- Bonferroni confidence intervals are simultaneous confidence intervals

$$\widetilde{\mathrm{CI}}_i = [y_i - z_{1-\alpha/2m}, y_i + z_{1-\alpha/2m}]$$

  Bonferroni confidence intervals are simple to use but conservative, i.e. $\mathrm{P}(\bigcap_{i=1}^{m} \{\widetilde{\mathrm{CI}}_i \ni \theta_i\}) > 1 - \alpha$

False coverage rate

Often researchers examine many parameters at once and report confidence intervals only for selected ones. Confidence intervals may have reduced coverage probability after selection. One quantity we may want to control is the false coverage rate.

Definition

The false coverage rate is defined as

$$\mathrm{FCR} = \mathbb{E}\Big(\frac{V_{\mathrm{CI}}}{R_{\mathrm{CI}} \vee 1}\Big)$$

where R_{CI} is the number of selected parameters and V_{CI} the number of constructed confidence intervals not covering

Note that without selection, the CIs control the FCR since

$$\mathbb{E}\Big(\frac{\sum_{i=1}^{m} \mathbb{1}\{\theta_i \notin \mathrm{CI}_i\}}{m}\Big) = \frac{\sum_{i=1}^{m} \mathbb{E}(\mathbb{1}\{\theta_i \notin \mathrm{CI}_i\})}{m} \leq \alpha$$

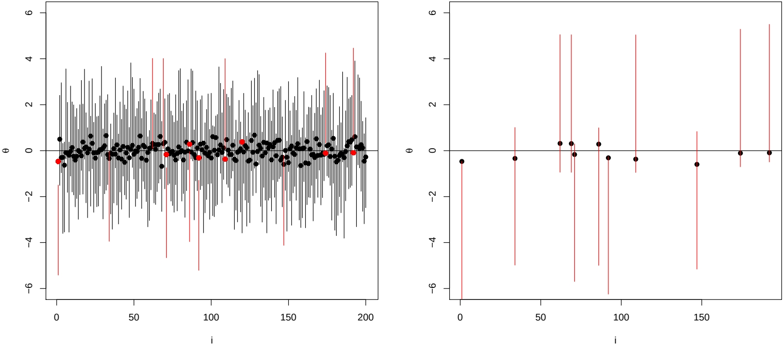
Benjamini & Yekutieli

- Let $y_1, \ldots, y_m$ independent with $y_i \sim N(\theta_i, 1)$
- Examples of selection rules (defined a priori):
  Unadjusted $\mathcal{S} = \{i \in \{1, \ldots, m\} : |y_i| \geq z_{1-\alpha/2}\}$
  Benjamini-Hochberg $\mathcal{S} = \{i \in \{1, \ldots, m\} : p_i \leq \frac{i^* \alpha}{m}\}$
  where $p_i = 2[1 - \Phi(|y_i|)]$ and $i^* = \max\{i : p_{(i)} \leq i\alpha/m\}$
- Benjamini & Yekutieli method
  1. Apply the selection rule and obtain $\mathcal{S}$
  2. Construct

$$\widetilde{\mathrm{CI}}_i = [y_i - z_{1-\tilde{\alpha}/2}, y_i + z_{1-\tilde{\alpha}/2}] \qquad i \in \mathcal{S}$$

  where $\tilde{\alpha} = \dfrac{\alpha|\mathcal{S}|}{m}$

  controls the FCR at α (assuming independence and simple selection rules)



$7/11 = 63.6\%$ selected CIs are covering, but 0% if adjusted