Notes: Computer Age Statistical Inference – Ch 15 Multiple Testing

Yingbo Li

04/08/2022

Table of Contents

Classical (Before Computer Age) Multiple Testing Corrections

- Bonferroni Correction
- Family-wise Error Rate

False Discovery Rates

- Benjamini-Hochberg FDR control
- An empirical Bayes view

Local False Discovery Rates

Empirical Null

Background and notations

- Before computer age, multiple testing may only involve 10 or 20 tests. With the emerge of biomedical (microarray) data, multiple testing may need to evaluate several thousands of tests

- Notations

  - N: total number of tests, e.g., number of genes.
  - z_i: the z-statistic of the i-th test. Note that if we perform tests other than z-test, say a t-test, then we can use inverse-cdf method to transform the t-statistic into a z-statistic, like below

    $$z_i = \Phi^{-1}\left[F_{df}(t_i)\right],$$

    where Φ is the standard normal cdf, and F is a t distribution cdf.
  - I_0: the indices of the true H_{0i}, having N_0 members. Usually, majority of hypotheses are null, so $\pi_0 = N_0/N$ is close to 1.

- Hypotheses: standard normal vs normal with a non-zero mean

  $$H_{0i} : z_i \sim \mathsf{N}(0,1) \longleftrightarrow H_{1i} : z_i \sim \mathsf{N}(\mu_i, 1)$$

  where μ_i is the effect size for test i

Example: the prostate data

- A microarray data of
  - $n = 102$ people, 52 prostate cancer patients and 50 normal controls
  - $N = 6033$ genes

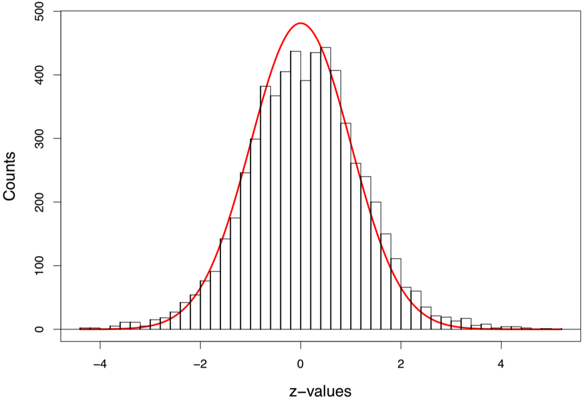


Figure 1: Histogram of 6033 z-values, with the scaled standard normal density curve in red

Classical multiple testing method 1: Bonferroni bound

- For an overall significance level α (usually $\alpha = 0.05$), with N simultaneous tests, the Bonferroni bound rejects the ith null hypothesis H_{0i} at individual significance level

$$p_i \leq \frac{\alpha}{N}$$

- Bonferroni bound is quite conservative!
  - For prostate data $N = 6033$ and $\alpha = 0.05$, the p-value rejection cutoff is very small: $p_i \leq 8.3 \times 10^{-6}$

Classical multiple testing method 2: FWER control

- The family-wise error rate is the probability of making even one false rejection

$$\text{FWER} = P(\text{reject any true } H_{0i})$$

- Bonferroni's procedure controls FWER, i.e., Bonferroni bound is more conservative than FWER control

$$\begin{aligned}\text{FWER} &= P\left\{\cup_{i\in I_0}\left(p_i \leq \frac{\alpha}{N}\right)\right\} \leq \sum_{i\in I_0} P\left(p_i \leq \frac{\alpha}{N}\right) \\ &= N_0\frac{\alpha}{N} \leq \alpha\end{aligned}$$

FWER control: Holm's procedure

1. Order the observed p-values from smallest to largest

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(i)} \ldots \leq p_{(N)}$$

2. Let $i_{\max}$ to be the largest index i such that

$$p_{(i)} \leq \text{Threshold(Holm's)} = \frac{\alpha}{N - i + 1}, \text{ for all } i \leq i_{\max}$$

3. Reject null hypotheses $H_{0(i)}$ for all $i \leq i_{\max}$

- FWER is usually still too conservative for large N, since it was originally developed for $N \leq 20$

An R function to implement Holm's procedure

```
## A function to obtain Holm's procedure p-value cutoff
## TO BE CORRECTED!
holm = function(pi, alpha=0.1){
  N = length(pi)
  idx = order(pi)
  reject = which(pi[idx] <= alpha/(N - 1:N + 1))

  return(idx[reject])
}
```

```
## Download prostate data's z-values
link = 'https://web.stanford.edu/~hastie/CASI_files/DATA/pro
prostz = c(read.table(link))$V1
## Convert to p-values
prostp = 1 - pnorm(prostz)
```

Illustrate Holm's procedure on the prostate data

```r
## Apply Holm's procedure on the prostate data
results = holm(prostp)
## Total number of rejected null hypotheses
r = length(results); r
```

```
## [1] 6
```

```r
## The largest z-value among non-rejected nulls
sort(prostz, decreasing = TRUE)[r + 1]
```

```
## [1] 4.13538
```

```r
## The smallest p-value among non-rejected nulls
sort(prostp)[r + 1]
```

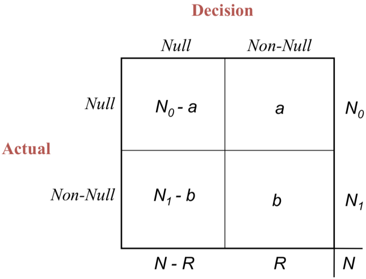
```
## [1] 1.771839e-05
```

False discovery proportion

- FDR control is a more liberal criterion (compared with FWER), thus it has become standard for large N multiple testing problems.
- False discovery proportion

$$\text{Fdp}(\mathcal{D}) = \begin{cases} a/R, & \text{if } R \neq 0 \\ 0, & \text{if } R = 0 \end{cases}$$

  - A decision rule $\mathcal{D}$ rejects R out of N null hypotheses
  - a of those are false discoveries (unobservable)



		Decision		
		Null	*Non-Null*	
Actual	*Null*	N_0 - a	a	N_0
	Non-Null	N_1 - b	b	N_1
		N - R	R	N

False discovery rate

- False discovery rates

$$\text{FDR}(\mathcal{D}) = E\{\text{Fdp}(\mathcal{D})\}$$

- A decision rule $\mathcal{D}$ controls FDR at level q, if

$$\text{FDR}(\mathcal{D}) \leq q$$

  - q is a prechosen value between 0 and 1

Benjamini-Hochberg FDR control

1. Order the observed p-values from smallest to largest

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(i)} \ldots \leq p_{(N)}$$

2. Let $i_{\max}$ to be the largest index i such that

$$p_{(i)} \leq \text{Threshold}(\mathcal{D}_q) = \frac{q}{N} i, \text{ for all } i \leq i_{\max}$$

3. Reject null hypotheses $H_{0(i)}$ for all $i \leq i_{\max}$

- Default choice $q = 0.1$
- Theorem: if the p-values are independent of each other, then the above procedure controls FDR at level q, i.e.,

$$\text{FDR}(\mathcal{D}_q) = \pi_0 q \leq q, \quad \text{where } \pi_0 = N_0/N$$

  - Usually, majority of the hypotheses are truly null, so π_0 is near 1

An R function to implement Benjamini-Hochberg FDR control

```r
## A function to obtain Holm's procedure p-value cutoff
## TO BE CORRECTED!
bh = function(pi, q=0.1){
  N = length(pi)
  idx = order(pi)
  reject = which(pi[idx] <= q/N * (1:N))

  return(idx[reject])
}
```

Illustrate Benjamini-Hochberg FDR control on the prostate data

```r
## Apply Holm's procedure on the prostate data
results = bh(prostp)
## Total number of rejected null hypotheses
r = length(results); r
```

```
## [1] 28
```

```r
## The largest z-value among non-rejected nulls
sort(prostz, decreasing = TRUE)[r + 1]
```

```
## [1] 3.293507
```

```r
## The smallest p-value among non-rejected nulls
sort(prostp)[r + 1]
```

```
## [1] 0.0004947302
```

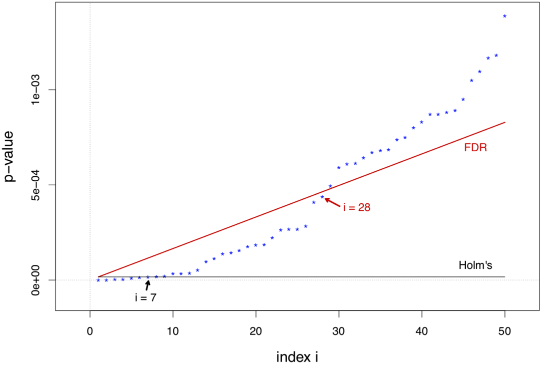
Comparing Holm's FWER control and Benjamini-Hochberg FDR control

- In the usual range of interest, large N and small i, the ratio

$$\frac{\text{Threshold}(\mathcal{D}_q)}{\text{Threshold(Holm's)}} = \frac{q}{\alpha}\left(1 - \frac{i-1}{N}\right) i$$

  increases with i almost linearly

- The figure below is about the prostate data, with $\alpha = q = 0.1$



Question about the FDR control procedure

1. Is controlling a rate (i.e., FDR) as meaningful as controlling a probability (of Type 1 error)?
2. How should q be chosen?
3. The control theorem depends on independence among the p-values. What if they're dependent, which is usually the case?
4. The FDR significance for one gene depends on the results of all other genes. Does this make sense?

Two-groups model

- Each of the N cases (e.g., genes) is
  - either null with prior probability π_0,
  - or non-null with probability $\pi_1 = 1 - \pi_0$
- For case i, its z-value z_i under H_{ij} for $j = 0, 1$ has density $f_j(z)$, cdf $F_j(z)$, and survival curve

$$S_j(z) = 1 - F_j(z)$$

- The mixture survival curve

$$S(z) = \pi_0 S_0(z) + \pi_1 S_1(z)$$

Bayesian false-discovery rate

- Suppose the observation z_i for case i is seen to exceed some threshold value z_0 (say $z_0 = 3$). By Bayes' rule, the Bayesian false-discovery rate is

$$\begin{aligned}\mathrm{Fdr}(z_0) &= P(\text{case } i \text{ is null} \mid z_i \geq z_0) \\ &= \frac{\pi_0 S_0(z_0)}{S(z_0)}\end{aligned}$$

- The "empirical" Bayes reflects in the estimation of the denominator: when N is large,

$$\hat{S}(z_0) = \frac{N(z_0)}{N}, \quad N(z_0) = \#\{z_i \geq z_0\}$$

- An empirical Bayes estimate of the Bayesian false-discovery rate

$$\widehat{\mathrm{Fdr}}(z_0) = \frac{\pi_0 S_0(z_0)}{\hat{S}(z_0)}$$

Connection between $\widehat{\text{Fdr}}$ and FDR controls

- Since $p_i = S_0(z_i)$ and $\hat{S}(z_{(i)}) = i/N$, the FDR control $\mathcal{D}_q$ algorithm

  $$p_{(i)} \leq \frac{i}{N} \cdot q$$

  becomes

  $$S_0(z_{(i)}) \leq \hat{S}(z_{(i)}) \cdot q,$$

  After rearranging the above formula, we have its Bayesian Fdr bounded

  $$\widehat{\text{Fdr}}(z_0) \leq \pi_0 q \tag{1}$$

- The FDR control algorithm is in fact rejecting those cases for which the empirical Bayes posterior probability of nullness is too small

Answer the 4 questions about the FDR control

1. (Rate vs probability) FDR control does relate to the posterior probability of nullness

2. (Choice of q) We can set q according to the maximum tolerable amount of Bayes risk of nullness, usually after taking $\pi_0 = 1$ in (1)

3. (Independence) Most often the z_i, and hence the p_i, are correlated. However even under correlation, $\hat{S}(z_0)$ is still an unbiased estimator for $S_(z_0)$, making $\widehat{\text{Fdr}}(z_0)$ nearly unbiased for $\text{Fdr}(z_0)$.
   - There is a price to be paid for correlation, which increases the *variance* of $\hat{S}(z_0)$ and $\widehat{\text{Fdr}}(z_0)$

4. (Rejecting one test depending on others) In the Bayes two-group model, the number of null cases z_i exceeding some threshold z_0 has *fixed* expectation $N\pi_0 S_0(z_0)$. So an increase in the number of z_i exceeding z_0 must come from a heavier right tail for $f_1(z)$, implying a greater posterior probability of non-nullness $\text{Fdr}(z_0)$.
   - This emphasizes the "learning from the experience of others"

Local false discovery rates

- Having observed test statistic z_i equal to some value z_0, we should be more interested in the probability of nullness given $z_i = z_0$ than $z_i \geq z_0$

- Local false discovery rate

$$\begin{aligned}\text{fdr}(z_0) &= P(\text{case } i \text{ is null} \mid z_i = z_0) \\ &= \frac{\pi_0 f_0(z_0)}{f(z_0)}\end{aligned}$$

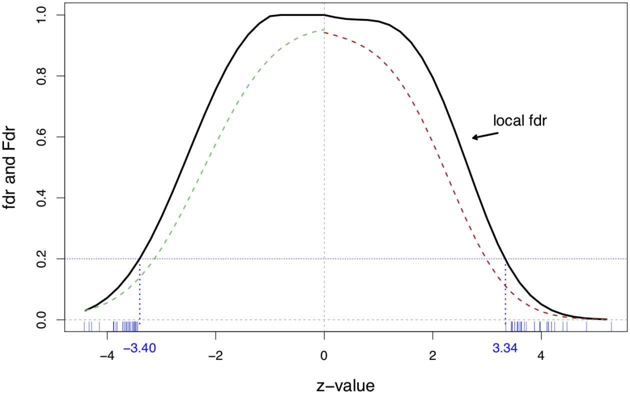
- After drawing a smooth curve $\hat{f}(z)$ through the histogram of the z-values, we get the estimate

$$\widehat{\text{fdr}}(z_0) = \frac{\pi_0 f_0(z_0)}{\hat{f}(z_0)}$$

  - the null proportion π_0 can either be estimated or set equal to 1

A fourth-degree log polynomial Poisson regression fit to the histogram, on the prostate data

- Solid line is the local $\widehat{\text{fdr}}(z)$ and dashed lines are tail-area $\widehat{\text{Fdr}}(z)$
- 27 genes on the right and 25 one the left have $\widehat{\text{fdr}}(z_i) \leq 0.2$



The default cutoff for local fdr

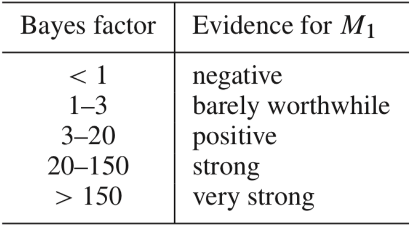
- The cutoff $\widehat{\text{fdr}}(z_i) \leq 0.2$ is equivalent to

$$\frac{f_1(z)}{f_0(z)} \geq 4\frac{\pi_0}{\pi_1}$$

- Assuming $\pi_0 \geq 0.9$, this makes the factor factor quite large

$$\frac{f_1(z)}{f_0(z)} \geq 36$$

  This is "strong evidence" against the null hypothesis in Jeffrey's scale of evidence for the interpretation of Bayes factors



Bayes factor	Evidence for M_1
< 1	negative
1–3	barely worthwhile
3–20	positive
20–150	strong
> 150	very strong

Relation between the local and tail-area fdr's

- Since

$$\mathsf{Fdr}(z_0) = E\left(\mathsf{fdr}(z) \mid z \geq z_0\right)$$

  Therefore

$$\mathsf{Fdr}(z_0) < \mathsf{fdr}(z_0)$$

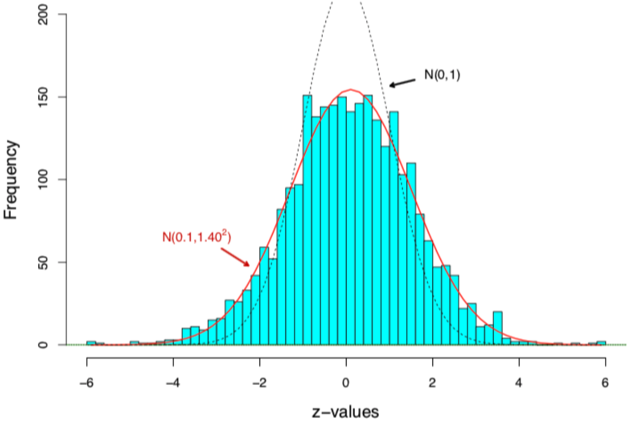
- Thus, the conventional significant cutoffs are

$$\widehat{\mathsf{Fdr}}(z) \leq 0.1$$

$$\widehat{\mathsf{fdr}}(z) \leq 0.2$$

Empirical null

- Large scale applications may allow us to empirically determine a more realistic null distribution than $H_{0i} : z_i \sim \mathrm{N}(0, 1)$
- In the police data, a $\mathrm{N}(0, 1)$ curve is too narrow for the null. Actually, an MLE fit to central data gives $\mathrm{N}(0.10, 1.40^2)$ as the empirical null



Empirical null estimation

- The theoretical null $z_i \sim \mathsf{N}(0,1)$ is not completely wrong, but needs adjustment for the dataset at hand
- Under the two-group model, with $f_0(z)$ normal but not necessarily standard normal

  $$f_0(z) \sim \mathsf{N}(\delta_0, \sigma_0^2),$$

  to compute the local fdr$(z) = \pi_0 f_0(z)/f(z)$, we need to estimate three parameters $(\delta_0, \sigma_0, \pi_0)$
- Our key assumption is that π_0 is large, say $\pi_0 \geq 0.9$, and most of the z_i near 0 are null.
- The algorithm `locfdr` begins by selecting a set $\mathcal{A}_0$ near $z = 0$ and assumes that all the z_i in $\mathcal{A}_0$ are null
- Maximum likelihood based on the numbers and values of z_i in $\mathcal{A}_0$ yield the empirical null estimates $(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0)$

References

- Efron, Bradley and Hastie, Trevor (2016), *Computer Age Statistical Inference*. Cambridge University Press
- Links to the prostate data
  - The 6033×102 data matrix: *prostmat.csv*
  - The 6033 z-values: *prostz.txt*
- A list of FDR methods in R: http://www.strimmerlab.org/notes/fdr.html