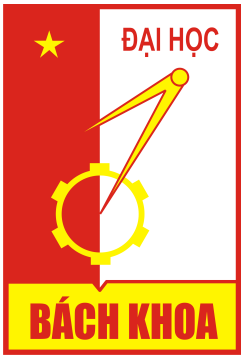
ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



BÀI TOÁN KIỂM ĐỊNH ĐA GIẢ THUYẾT VÀ ỨNG DỤNG

ĐỒ ÁN I

Chuyên ngành: Toán Tin

Chuyên sâu: Toán ứng dụng

Chữ ký GVHD

Giảng viên hướng dẫn:	**TS. Nguyễn Văn Hạnh**
Sinh viên thực hiện:	**Nguyễn Lương Quỳnh Trang**
MSSV:	**20206175**
Lớp:	**Toán Tin 03 - K65**

Hà Nội, Ngày 1 tháng 6 năm 2024

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đồ án

(a) Mục tiêu: Tìm hiểu về kiểm định đa giả thuyết và một số phương pháp hiệu chỉnh, kiểm soát sai số cho bài toán kiểm định đa giả thuyết; ứng dụng các phương pháp đã tìm hiểu vào kiểm soát sai số đối với bài toán thực tế.

(b) Nội dung: Giới thiệu về kiểm định đa giả thuyết và một số phương pháp hiệu chỉnh, kiểm soát sai số; thực hiện các phương pháp với bộ dữ liệu có sẵn.

2. Kết quả đạt được

(a) Các khái niệm cơ bản về bài toán kiểm định đa giả thuyết.

(b) Một số phương pháp hiệu chỉnh để kiểm soát sai số.

(c) Kiểm soát sai số cho bài toán kiểm định đa giả thuyết trong thực tế.

3. Ý thức làm việc của sinh viên

(a)

(b)

Hà Nội, ngày 27 tháng 7 năm 2023

Giảng viên hướng dẫn

TS. Nguyễn Văn Hạnh

Mở đầu

Trong thống kê, kiểm định giả thuyết là phương pháp suy luận, sử dụng dữ liệu từ thí nghiệm hoặc từ quan sát các mẫu trong thực tế, từ đó thông qua một số bước tính toán để thu được một kết quả, đưa ra quyết định hay kết luận về dữ liệu đó. Một kết luận được gọi là có thể tin tưởng được theo thống kê nếu xác suất xảy ra kết luận đó đủ lớn ở một mức nhất định.

Bài toán kiểm định đa giả thuyết, hay còn gọi là kiểm định bội, xuất hiện khi để đưa ra một kết luận, người ta xem xét một bài toán đó thông qua một tập các vấn đề con, mà mỗi vấn đề con đó tương ứng với một bài toán kiểm định giả thuyết đơn giản. Kết luận dựa trên càng nhiều giả thuyết đơn thì xác suất xảy ra sai lầm càng cao. Các nhà thống kê đã quan tâm đến vấn đề này và nghiên cứu để đưa ra nhiều phương pháp nhằm kiểm soát sai số toàn cục, nâng cao mức độ tin cậy của kết luận cuối cùng.

Báo cáo này trình bày những khái niệm cơ bản về bài toán kiểm định đa giả thuyết, nêu ra những hướng tiếp cận và phương pháp tiêu biểu theo một số hướng tiếp cận đó để kiểm soát sai số toàn cục của bài toán kiểm định đa giả thuyết. Để có cái nhìn trực quan hơn về các phương pháp này, báo cáo cũng ứng dụng các phương pháp này để thực hiện với bộ dữ liệu thực tế.

Cấu trúc của Đồ án

1. Giới thiệu bài toán
2. Một số phương pháp hiệu chỉnh để kiểm soát sai số
3. Ứng dụng
4. Kết luận

Một số ký hiệu và chữ viết tắt

Trong báo cáo sử dụng một số thuật ngữ viết tắt và ký hiệu như sau:

$FWER$	Family-wise error rate Xác suất xảy ra ít nhất một sai lầm loại 1
FDR	False discovery rate Tỷ lệ phát hiện sai lầm
$\mathbb{P}(A)$	Xác suất xảy ra sự kiện A
$\mathbb{E}(V)$	Trung bình của biến ngẫu nhiên V
α	Mức ý nghĩa của kiểm định giả thuyết, thường được lấy bằng 0.05
H_0	Giả thuyết trong một kiểm định thống kê, ngược lại gọi là đối thuyết
H_i	Giả thuyết của kiểm định thống kê thứ i

Dù đã cố gắng để hoàn thiện nhưng khó tránh khỏi những sai sót, em rất mong nhận được lời nhận xét, góp ý của các thầy cô để báo cáo này có thể được hoàn chỉnh hơn nữa.

Lời cảm ơn

Trong quá trình tìm hiểu đề tài và hoàn thiện báo cáo, em xin gửi lời cảm ơn sâu sắc đến giảng viên hướng dẫn, TS. Nguyễn Văn Hạnh đã cho em gợi ý và định hướng về đề tài thú vị này, đồng thời cũng là người trực tiếp chỉ dẫn, giải đáp thắc mắc và cho em những góp ý để em hiểu rõ hơn về đề tài và có thể hoàn thành báo cáo một cách tốt nhất.

Em xin chân thành cảm ơn!

Hà Nội, ngày 27 tháng 7 năm 2023

Tác giả đồ án

Nguyễn Lương Quỳnh Trang

Mục lục

Danh sách hình vẽ

Chương 1

Giới thiệu bài toán

1.1 Kiểm định giả thuyết và các loại sai lầm

Một kiểm định giả thuyết tập trung vào một cặp giả thuyết - đối thuyết, ở đây ta gọi giả thuyết là H_0. Trong trường hợp này, ta dựa vào một kết quả thu được thông qua một kiểm định thống kê để quyết định xem có thể chấp nhận giả thuyết được đưa ra hay bác bỏ giả thuyết đó (tức là chấp nhận đối thuyết). Sai lầm thống kê thể hiện xác suất đưa ra một quyết định không chính xác.

Các loại sai lầm thống kê được minh họa như trong bảng sau:

Quyết định	H_0 đúng	H_0 không đúng
Chấp nhận H_0	Quyết định đúng $(1-\alpha)$	Sai lầm loại 2 (β)
Bác bỏ H_0	Sai lầm loại 1 (α)	Quyết định đúng $(1-\beta)$

Hình 1.1: Các loại sai lầm thống kê

Bác bỏ giả thuyết trong khi nó thật sự đúng được gọi là sai lầm loại 1, hay dương tính giả. Chấp nhận giả thuyết khi nó thật ra không đúng được gọi là sai lầm loại 2, hay âm tính giả. Tỷ lệ sai lầm loại 1 thường được ký hiệu là α, trong khi đó tỷ lệ sai lầm loại 2 thường được ký hiệu là β.

Khi nghiên cứu một bài toán kiểm định giả thuyết, ta thường mong muốn

kiểm soát tỷ lệ dương tính giả α, trong khi vẫn duy trì lực lượng kiểm định $1-\beta$ đủ tốt. α thường được gọi là mức ý nghĩa.

1.2 Kiểm định đa giả thuyết

Lấy một ví dụ, micro-array là một kỹ thuật điển hình để kiểm tra mức độ biểu hiện của các gen trong lĩnh vực Y - Sinh học. Khi kiểm tra mức độ biểu hiện của bộ gen, ta cần kiểm tra thông tin về nhiều gen riêng biệt, đồng nghĩa với nhiều giả thuyết cần phải kiểm định. Trong một thí nghiệm micro-array, ta cần thực hiện việc quan sát trên khoảng 1000 gen ở 2 trạng thái người thường - người bệnh, để xác định xem các gen này có biểu hiện bất thường hay không, tức là cần thực hiện kiểm định 1000 cặp giả thuyết - đối thuyết riêng biệt. Nếu sử dụng *p-giá trị* với mức ý nghĩa tiêu chuẩn là 0.05, ta biết rằng có ngẫu nhiên khoảng 50 gen được coi là "đáng chú ý", tức là 50 gen này thực chất không có biểu hiện bất thường nhưng qua kiểm định lại nhận được kết quả là bất thường.

Một cách tổng quát, nếu ta thực hiện kiểm định m giả thuyết, với mức ý nghĩa của mỗi giả thuyết là α. Khi này tỷ lệ có ít nhất một cặp giả thuyết rơi vào sai lầm loại 1 là bao nhiêu? Ta thử thực hiện một vài phép tính sau:

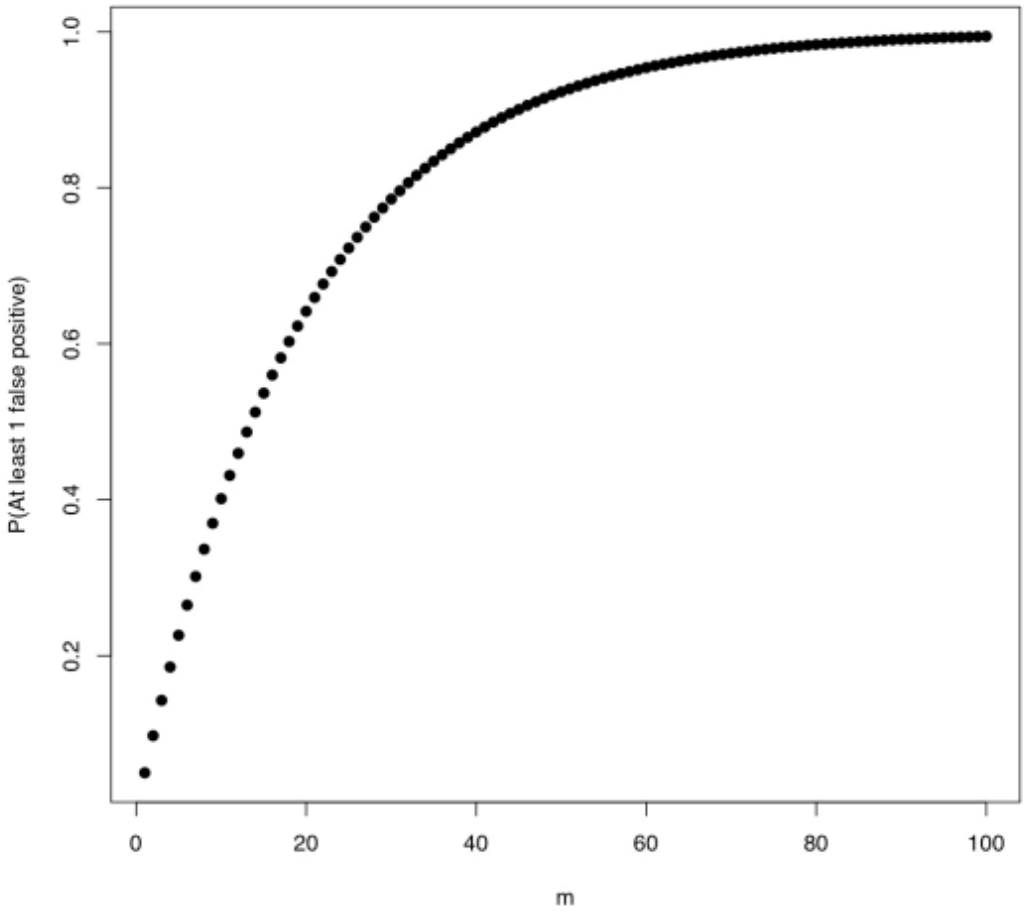
$$\mathbb{P}(\text{xảy ra 1 sai lầm}) = \alpha$$

$$\mathbb{P}(\text{không sai lầm trong 1 giả thuyết}) = 1-\alpha$$

$$\mathbb{P}(\text{không sai lầm trong cả } m \text{ giả thuyết}) = (1-\alpha)^m$$

$$\mathbb{P}(\text{xảy ra ít nhất 1 sai lầm trong } m \text{ giả thuyết}) = 1-(1-\alpha)^m$$

Như vậy, với số lượng cặp giả thiết càng lớn thì tỷ lệ xảy ra ít nhất một sai lầm loại 1 là lớn dần và tiến dần về 1. Nhìn vào đồ thị minh họa sau đây, ta thấy chỉ cần thực hiện kiểm định đồng thời với khoảng ít nhất 100 giả thuyết thì xác suất xảy ra ít nhất một sai lầm loại 1 gần như là bằng 1, tức là ta gần như chắc chắn sẽ mắc phải sai lầm.



Hình 1.2: Xác suất xảy ra ít nhất một sai lầm loại 1 với số giả thuyết tương ứng

Do đó, các nhà thống kê nghiên cứu những phương pháp khác nhau nhằm kiểm soát tỷ lệ xảy ra sai lầm loại 1 trong các bài toán kiểm định đa giả thuyết. Đây là một lĩnh vực thiết thực của thống kê, với nhiều phương pháp khác nhau đã được đưa ra. Mặc dù các phương pháp này hướng tới cùng một mục đích, nhưung hướng tiếp cận về cơ bản là khác nhau.

Điểm khác nhau trong các phương án tiếp cận ở đây là độ đo tỷ lệ xuất hiện sai lầm loại 1. Trước khi giới thiệu về các độ đo, ta thống nhất một số ký hiệu về đếm số sai lầm như sau:

Giả sử ta kiểm định m giả thuyết, gọi m_0 là số lượng giả thuyết đúng, $R_{\mathcal{D}}$ là số lượng giả thuyết bị bác bỏ, $F_{\mathcal{D}}$ là số lượng sai lầm loại 1.

Quyết định	H_0 đúng	H_0 không đúng	Tổng
Chấp nhận H_0	U	T	$m - R_{\mathcal{D}}$
Bác bỏ H_0	$F_{\mathcal{D}}$	S	$R_{\mathcal{D}}$
	m_0	$m - m_0$	

Hình 1.3: Quy ước về ký hiệu số lượng giả thuyết

Quay lại với các hướng tiếp cận đã nhắc đến ở trên, một số độ đo được sử dụng là:

- Trung bình của số sai lầm loại 1 trên tổng số giả thuyết (Per comparison error rate):
$$PCER = \frac{\mathbb{E}(F_{\mathcal{D}})}{m}$$
- Trung bình của số sai lầm loại 1 (Per-family error rate):
$$PFER = \mathbb{E}(F_{\mathcal{D}})$$
- Xác suất xuất hiện ít nhất một sai lầm loại 1 (Family-wise error rate):
$$FWER = \mathbb{P}(F_{\mathcal{D}} \geq 1)$$
- Tỷ lệ phát hiện sai lầm (False discovery rate):
$$FDR = \mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\middle| R_{\mathcal{D}} > 0\right)\mathbb{P}(R_{\mathcal{D}} > 0)$$
- Tỷ lệ phát hiện sai lầm dương (Positive false discovery rate):
$$pFDR = \mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\middle| R_{\mathcal{D}} > 0\right)$$

Trong báo cáo này, ta quan tâm đến 2 độ đo là $FWER$ và FDR.

1.3 Bài toán

Để có một cái nhìn tổng quan và thống nhất xuyên suốt báo cáo theo các phương pháp sẽ được trình bày ở sau, phần này sẽ nêu tóm tắt bài toán tổng quát cùng các ký hiệu liên quan.

Giả sử ta kiểm định m giả thuyết, các giả thuyết được ký hiệu lần lượt là H_1, $H_2, \ldots$, H_m. $\mathcal{K} = \{1, 2, \ldots . m\}$ là tập các chỉ số của giả thuyết. $\mathcal{N} \subseteq \mathcal{K}$ là tập các chỉ số của các giả thuyết đúng. $m_0 = |\mathcal{N}|$ là số giả thuyết đúng. P là một biến ngẫu nhiên thể hiện cho *p-giá trị*. Với mỗi $i \in \mathcal{K}$, giả thuyết H_i tương ứng với một *p-giá trị* p_i, là giá trị thực tế của biến ngẫu nhiên P_i. $\mathcal{D} \subseteq 2^{\mathcal{K}}$ là tập các giả thuyết H_i bị bác bỏ. $R_{\mathcal{D}} = |\mathcal{D}|$ là số giả thuyết bị bác bỏ. Như vậy, ta có $\mathcal{D} \cap \mathcal{N}$ là tập các giả thuyết bị bác bỏ sai. Gọi $F_{\mathcal{D}} = |\mathcal{D} \cap \mathcal{N}|$, $F_{\mathcal{D}}$ chính là số giả thuyết bị bác bỏ sai, hay số sai lầm loại 1.

Sau khi sắp xếp các *p-giá trị* theo thứ tự từ nhỏ đến lớn, ta ký hiệu các *p-giá trị* đã được sắp xếp là $p_{(i)}$, với giả thuyết tương ứng là $H_{(i)}$.

Khi kiểm định từng giả thuyết riêng lẻ với mức ý nghĩa α, ta bác bỏ giả thuyết H_i khi $p_i \leq \alpha$, ngược lại, chấp nhận H_i khi $p_i > \alpha$. Tuy nhiên, làm như vậy sẽ gây ra tỷ lệ sai lầm lớn khi ta muốn xem xét đồng thời các giả thuyết H_i, như đã nêu ở trên.

Trong phần sau, báo cáo trình bày về 2 độ đo là $FWER$ và FDR, và các phương pháp tiêu biểu hiệu chỉnh dựa trên việc kiểm soát 2 chỉ số này, khi xem xét đồng thời m giả thuyết H_i, để đảm bảo tỷ lệ sai lầm tổng thể không vượt quá mức α nhất định.

Chương 2

Một số phương pháp hiệu chỉnh sai số cho bài toán kiểm định đa giả thuyết

2.1 Hiệu chỉnh sai số $FWER$

$FWER$ (Family-wise error rate) là xác suất xảy ra ít nhất một sai lầm loại 1 trong số tất cả các giả thuyết được kiểm định đồng thời:

$$FWER = \mathbb{P}(F_{\mathcal{D}} \geq 1)$$

Có nhiều quy trình được đưa ra để kiểm soát $FWER$, được phân loại thành 2 hướng chính:

- Hiệu chỉnh riêng lẻ (single-step): Hiệu chỉnh riêng lẻ các *p-giá trị* theo cùng một tỷ lệ tương đương nhau.
- Hiệu chỉnh lần lượt (sequential): Sắp thứ tự các *p-giá trị* rồi hiệu chỉnh ở các mức độ khác nhau cho các kiểm định khác nhau. Tiêu chí để sắp xếp mức độ cho các *p-giá trị* được xác định linh hoạt tùy vào từng trường hợp hay từng phương pháp được sử dụng.

2.1.1 Phương pháp Bonferroni

Bonferroni là một phương pháp đơn giản để kiểm soát sai số $FWER$ bằng cách hiệu chỉnh riêng lẻ *p-giá trị*. Nó đảm bảo rằng $FWER$ duy trì ở mức không vượt quá α khi thực hiện m kiểm định độc lập.

Bonferroni xác định một ngưỡng tin cậy mới cho từng kiểm định trong số m kiểm định được thực hiện, bằng cách chia α cho số kiểm định là m. Tức là ta sẽ so sánh *p-giá trị* của từng kiểm định với mức ý nghĩa đã được hiệu chỉnh là $\frac{\alpha}{m}$.

Hay nói cách khác, Bonferroni hiệu chỉnh *p-giá trị* của từng kiểm định bằng cách nhân giá trị đó với m. Nếu *p-giá trị* sau khi hiệu chỉnh lớn hơn 1, ta sẽ thay thế nó bằng 1.

$$\tilde{p}_i = \min[mp_i, 1]$$

$\tilde{p}_i$ là một *p-giá trị* đã hiệu chỉnh ứng với giả thuyết H_i. Ta bác bỏ các giả thuyết H_i nếu có $\tilde{p}_i \leq \alpha$.

Bổ đề 2.1. *Khi giả thuyết H_0 là đúng, p-giá trị theo một tiêu chuẩn thống kê tuân theo phân phối đều liên tục:*

$$p \sim \mathcal{U}(0, 1).$$

Chứng minh. Không mất tính tổng quát, ta xét với một kiểm định giả thuyết hai phía.

Khi đó, *p-giá trị* là một hàm theo tiêu chuẩn thống kê T:

$$P = 2. \min\left(F_T(T), 1 - F_T(T)\right)$$

với $F_T(T)$ là hàm phân phối tích lũy của của tiêu chuẩn thống kê T khi H_0 đúng. Tức là, tại một giá trị quan sát t_{qs} cụ thể, ta có:

$$p = 2. \min\left(F_T(t_{qs}), 1 - F_T(t_{qs})\right)$$

Hàm phân phối tích lũy cho *p-giá trị* là:

- Với $\min\left(F_T(t_{qs}), 1 - F_T(t_{qs})\right) = F_T(t_{qs})$, ta có:

$$\begin{aligned} F_P(p) &= \mathbb{P}(P < p) \\ &= \mathbb{P}(F_T(T) < p) \\ &= \mathbb{P}(T < F_T^{-1}(p)) \\ &= F_T(F_T^{-1}(p)) \\ &= p \end{aligned}$$

- Tương tự với $\min\left(F_T(t_{qs}), 1 - F_T(t_{qs})\right) = 1 - F_T(t_{qs})$, ta có:

$$\begin{aligned} F_P(p) &= \mathbb{P}(P < p) \\ &= \mathbb{P}(F_T^{-1}(P) < F_T^{-1}(p)) \\ &= F_T(F_T^{-1}(p)) \\ &= p \end{aligned}$$

Ta thấy, hàm phân phối tích lũy cho *p-giá trị* là hàm phân phối tích lũy của một phân phối đều liên lục trên khoảng $[0, 1]$.

□

Khi đó, $\mathbb{P}(P < p) = F_P(p) = p$.

Định lý 2.1. *Quy trình Bonferroni nêu trên luôn kiểm soát sai số FWER ở mức không vượt quá mức ý nghĩa α đã định trước.*

Chứng minh. Ta có m giả thuyết được kiểm định là $H_1, H_2, \ldots, H_m$, với m_0 là số giả thuyết thật sự đúng, trong đó $F_{\mathcal{D}}$ là số giả thuyết đúng nhưng bị bác bỏ, tức là sai lầm loại 1.

$\mathcal{N}$ là tập các chỉ số i sao cho H_i đúng:

$$\mathcal{N} = \{i | H_i \text{ đúng}\}, \quad m_0 = |\mathcal{N}|$$

Gọi A_i là sự kiện H_i bị bác bỏ sai, tức là $p_i \leq \frac{\alpha}{m}$ với điều kiện H_i đúng.
Ta có:

$$FWER = \mathbb{P}(F_{\mathcal{D}} \geq 1) = \mathbb{P}\left(\bigcup_{i \in I} A_i\right) \leq \sum_{i \in I} \mathbb{P}(A_i) \leq m_0 . \frac{\alpha}{m} \leq \alpha.$$

□

2.1.2 Phương pháp Holm

Phương pháp Holm là phương pháp hiệu chỉnh lần lượt để kiểm soát $FWER$ đơn giản nhất.

Với m giả thuyết kiểm định đồng thời, trước tiên sắp xếp các *p-giá trị* thu được theo thứ tự không giảm $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$, tương ứng với các giả thuyết $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. Gọi i là hạng của $p_{(i)}$.

Để kiểm soát $FWER$ ở mức ý nghĩa α, các *p-giá trị* được hiệu chỉnh theo công thức:

$$\tilde{p}_{(i)} = \min[(m + 1 - i).p_{(i)}, 1]$$

Ở đây, ta không nhân tất cả *p-giá trị* với cùng một nhân tử m, mà các nhân tử sẽ được tính tương ứng với hạng của *p-giá trị*. Nếu giá trị của *p-giá trị* hiệu chỉnh lớn hơn 1 thì ta thay thế nó bằng 1.

Sau khi hiệu chỉnh, ta có các *p-giá trị* hiệu chỉnh $\tilde{p}_{(i)}$ ứng với giả thuyết $H_{(i)}$. Bác bỏ lần lượt các H_i có $\tilde{p}_{(i)} \leq \alpha$ từ $i = 1$ về sau. Với k là số nhỏ nhất thỏa mãn $\tilde{p}_{(k)} > \alpha$, ta chấp nhận tất cả $H_{(i)}$ với $i \geq k$.

Định lý 2.2. *Quy trình Holm nêu trên luôn kiểm soát sai số $FWER$ ở mức không vượt quá mức ý nghĩa α đã định trước.*

Chứng minh. Ta có m giả thuyết được kiểm định là $H_1, H_2, \ldots, H_m$, sắp xếp lại theo thứ tự *p-giá trị* không giảm là $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$, với m_0 là số giả thuyết thật sự đúng, trong đó $F_{\mathcal{D}}$ là số giả thuyết đúng nhưng bị bác bỏ, tức là sai lầm

loại 1.

$\mathcal{N}$ là tập các chỉ số i sao cho $H_{(i)}$ đúng:

$$\mathcal{N} = \{i | H_{(i)} \text{ đúng}\}, \quad m_0 = |\mathcal{N}|$$

Gọi $A_{(i)}$ là sự kiện $H_{(i)}$ bị bác bỏ sai, tức là $p_{(i)} \leq \dfrac{\alpha}{m+1-i}$ với điều kiện $H_{(i)}$ đúng.

Giả sử ta bác bỏ ít nhất một trong số m_0 giả thuyết đúng. Gọi $H_{(k)}$ là giả thuyết đầu tiên bị bác bỏ sai. Như vậy, có ít nhất $k-1$ giả thuyết được bác bỏ đúng, là $H_{(1)}, H_{(2)}, \ldots, H_{(k-1)}$.

Ta có: $m_0 \leq m-(k-1) \Rightarrow \dfrac{1}{m_0} \geq \dfrac{1}{m+1-k} \Rightarrow p_{(k)} \leq \dfrac{\alpha}{m+1-k} \leq \dfrac{\alpha}{m_0}$

$\Rightarrow p_{(i)} \leq \dfrac{\alpha}{m+1-i} \leq \dfrac{\alpha}{m_0} \quad \forall i \geq k.$

Vậy:

$$FWER = \mathbb{P}(F_{\mathcal{D}} \geq 1) = \mathbb{P}\left(\bigcup_{i \in I} A_{(i)}\right) \leq \sum_{i \in I} \mathbb{P}(A_{(i)}) \leq m_0 . \frac{\alpha}{m_0} \leq \alpha.$$

□

2.1.3 Nhận xét

- Các phương pháp kiểm soát $FWER$ đảm bảo chắc chắn tỷ lệ xảy ra ít nhất một sai lầm loại 1 không vượt quá một ngưỡng nhất định. Do đó giảm được tỷ lệ sai lầm loại 1. Tuy nhiên, điều này khiến cho tỷ lệ sai lầm loại 2 tăng lên rất lớn, do có nhiều giả thuyết có giá trị *p-giá trị* tuy đã đủ nhỏ nhưng vẫn không đạt đến ngưỡng bị bác bỏ, từ đó làm giảm lực lượng thống kê.

- Đa số trường hợp, ta không cần kiểm soát $FWER$ một cách cực đoan như thế. Thực tế, ta có thể chấp nhận một số lượng nhất định những sai lầm loại 1 xảy ra.

- Kiểm soát $FWER$ hữu ích khi kết luận của kiểm định bội có xu hướng sai lầm khi ít nhất một giả thuyết trong số các giả thuyết đơn bị kết luận sai.

Do đó trong nhiều trường hợp người ta dùng đến độ đo FDR (false discovery rate).

2.2 Hiệu chỉnh FDR

Trong nhiều bài toán kiểm định bội, người ta quan tâm đến tỷ lệ số giả thuyết bị bác bỏ sai lầm, hay còn gọi là phát hiện sai lầm, trên tổng số giả thuyết bị bác bỏ nhiều hơn là số lượng các bác bỏ sai lầm. Trong trường hợp này, những mất mát (về chi phí và tài nguyên,...) phát sinh do bác bỏ nhầm tỷ lệ nghịch với số giả thuyết bị bác bỏ. Từ đó các nhà nghiên cứu mong muốn có thể kiểm soát trung bình của tỷ lệ này, gọi là tỷ lệ phát hiện sai lầm FDR (false discovery rate).

2.2.1 Định nghĩa tỷ lệ phát hiện sai lầm FDR

Tỷ lệ phát hiện sai lầm được quan sát qua biến ngẫu nhiên $Q = \dfrac{F_{\mathcal{D}}}{F_{\mathcal{D}} + S} = \dfrac{F_{\mathcal{D}}}{R_{\mathcal{D}}}$. Nếu không có giả thuyết nào bị bác bỏ, tức là $R_{\mathcal{D}} = 0$, hiển nhiên ta định nghĩa $Q = 0$. Tỷ lệ phát hiện sai lầm FDR được định nghĩa là trung bình của biến ngẫu nhiên Q:

$$FDR = \mathbb{E}(Q) = \mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right)$$

Tỷ lệ này có 2 tính chất quan trọng:

- Nếu thật sự tất cả các giả thuyết kiểm tra đều đúng, tức là $m_0 = m$ thì $FDR = FWER$. Trong trường hợp này, $S = 0$ và $F_{\mathcal{D}} = R_{\mathcal{D}}$. Do đó nếu $F_{\mathcal{D}} = 0$ thì $Q = 0$, nếu $F_{\mathcal{D}} > 0$ thì $Q = 1$, dẫn đến $\mathbb{P}(F_{\mathcal{D}} \leq 1) = \mathbb{E}(Q)$. Trường hợp kiểm soát FDR này còn được gọi là kiểm soát $FWER$ yếu.
- Nếu có ít nhất một giả thuyết sai, tức là $m_0 \leq m$ thì $FDR \leq FWER$.

Do đó các quy trình kiểm soát $FWER$ đều kiểm soát FDR. Tuy nhiên các quy trình kiểm soát FDR thì linh hoạt hơn và có thể thu được lực lượng thống

kê như mong đợi. Trong thực tế, số giả thuyết không đúng càng nhiều thì số dương tính thật S có xu hướng càng lớn, tức là lực lượng thống kê càng lớn.

2.2.2 Phương pháp Benjamini - Hochberg (BH)

Kiểm soát được biến ngẫu nhiên Q là một điều đáng mong đợi, tuy nhiên việc này là bất khả thi. Vấn đề tương tự xảy ra khi ta cố gắng kiểm soát $\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\middle| R_{\mathcal{D}} > 0\right)$. Benjamini và Hochberg đã đưa ra một công thức thay thế mà việc kiểm soát tỷ lệ này là khả thi:

$$FDR = \mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\middle| R_{\mathcal{D}} > 0\right)\mathbb{P}(R_{\mathcal{D}} > 0)$$

Thực hiện kiểm định m giả thuyết $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ với các *p-giá trị* tương ứng đã được sắp xếp theo thứ tự không giảm là $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.

Với k là số i lớn nhất thỏa mãn $p_{(i)} \leq \frac{i}{m}\alpha$,

$$\text{bác bỏ tất cả các giả thuyết } H_{(i)}, \quad i = 1, \ldots, k.$$

Định lý 2.3. *Với các kiểm định thống kê độc lập có số giả thuyết sai bất kỳ, quy trình trên luôn kiểm soát sai số FDR ở mức α được định trước.*

Chứng minh. Đặt $\alpha_r = \frac{\alpha.r}{m}$

$$FDR = \mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right) = \mathbb{E}\left(\frac{\sum\limits_{i\in\mathcal{N}} \mathbb{1}_{\{p_{(i)}\leq\alpha_{R_{\mathcal{D}}}\}}}{R_{\mathcal{D}}}\right) = \sum_{i\in\mathcal{N}}\sum_{r=1}^{m}\frac{1}{r}\mathbb{E}\left(\mathbb{1}_{\{p_{(i)}\leq\alpha_r\}}\mathbb{1}_{\{R_{\mathcal{D}}=r\}}\right) \quad (*)$$

Với $i \in \mathcal{N}$, gọi R_i là số giả thuyết bị bác bỏ nếu thay $P_{(i)} = 0$.

Do $\{P_{(i)} \leq \alpha_r, R_{\mathcal{D}} = r\} = \{P_{(i)} \leq \alpha_r, R_i = r\}$ với mọi i, r, ta có:

$$\mathbb{E}\left(\mathbb{1}_{\{p_{(i)}\leq\alpha_r\}}\mathbb{1}_{\{R_{\mathcal{D}}=r\}}\right) = \mathbb{E}\left(\mathbb{1}_{\{p_{(i)}\leq\alpha_r\}}\mathbb{1}_{\{R_i=r\}}\right)$$

Theo giả thiết $P_{(i)}$ độc lập với $P_{(j)}$, $j \in \mathcal{K} \setminus \{i\}$, nghĩa là $P_{(i)}$ độc lập với R_i. Với $i \in \mathcal{N}$:

$$\mathbb{E}\left(\mathbb{1}_{\{p_{(i)}\leq\alpha_r\}}\mathbb{1}_{\{R_i=r\}}\right) = \mathbb{P}(P_{(i)} \leq \alpha_r)\mathbb{P}(R_j = r) = \frac{\alpha r}{K}\mathbb{P}(R_i = r)$$

Thay vào (*) ta có:

$$\mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right) = \frac{\alpha}{m}\sum_{i\in\mathcal{N}}\sum_{r=1}^{m}\mathbb{P}(R_i = r) = \frac{m_0}{m}\alpha.$$

□

2.2.3 Phương pháp *e*-BH kiểm soát FDR với *e-giá trị*

Đây là một phương pháp tương tự như BH nhưng được dùng cho *e-giá trị*, nên nó được gọi là e-BH. Đầu vào gồm 3 thành phần:

a. m *e-giá trị* $e_1, e_2, \ldots, e_m$ ứng với m giả thuyết là $H_1, H_2, \ldots, H_m$

b. Mức FDR α

c. (Không bắt buộc) Thông tin phân phối hoặc giả định về *e-giá trị*

Trong khi a. và b. là 2 thành phần đầu vào tương tự như ở quy trình BH, thì c. là một điểm mới của quy trình e-BH.

Quy trình e-BH có thể được mô tả ngắn gọn trong 2 bước như sau:

1. (Không bắt buộc) Hiệu chỉnh *e-giá trị* ban đầu sử dụng thông tin ở c.

2. Thực hiện thuật toán e-BH cơ sở với các *e-giá trị* đã hiệu chỉnh và mức α.

Quy trình *e*-BH cơ sở

Gọi $e'_1, e'_2, \ldots, e'_m$ là các *e-giá trị* đã hiệu chỉnh thu được ở bước 1. của quy trình. Gọi $e'_{(i)}$ là giá trị thứ i của $e'_1, e'_2, \ldots, e'_m$ đã được sắp xếp từ lớn nhất đến nhỏ nhất, khi đó $e'_{(1)}$ là *e-giá trị* đã hiệu chỉnh lớn nhất.

Với k là số i lớn nhất thỏa mãn $\dfrac{ie'_{(i)}}{m} \geq \dfrac{1}{\alpha}$,

$$\text{bác bỏ tất cả các giả thuyết } H_i, \quad i = 1, \ldots, k.$$

Nhận xét: Quy trình e-BH tương tự như quy trình BH áp dụng cho $(e_1^{-1}, e_2^{-1}, \ldots, e_m^{-1})$.

Định lý 2.4. *Với e-giá trị tùy ý, thủ tục e-BH với mức $\alpha \in (0,1)$ có FDR được kiểm soát không vượt quá $\alpha \frac{m_0}{m}$.*

Chứng minh. Theo định nghĩa ta có:

$$E_{(i)} \geq \frac{m}{\alpha R_{\mathcal{D}}}$$

Do đó:

$$\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}} = \frac{|\mathcal{D} \cap \mathcal{N}|}{R_{\mathcal{D}} \vee 1} = \sum_{i \in \mathcal{N}} \frac{\mathbb{1}_{\{i \in \mathcal{D}\}}}{R_{\mathcal{D}} \vee 1} \leq \sum_{i \in \mathcal{N}} \frac{\mathbb{1}_{\{i \in \mathcal{D}\}} \alpha E_{(i)}}{m} \leq \sum_{k \in \mathcal{N}} \frac{\alpha E_{(i)}}{m}$$

Do $\mathbb{E}(E_{(i)}) \leq 1$ với $i \in \mathcal{N}$, ta có:

$$\mathbb{E}\left(\frac{F_{\mathcal{D}}}{R_{\mathcal{D}}}\right) \leq \mathbb{E}\left(\sum_{i \in \mathcal{N}} \frac{\alpha E_{(i)}}{m}\right) \leq \frac{\alpha m_0}{m}.$$

□

Hiệu chỉnh *e-giá trị* Để hiệu chỉnh *e-giá trị*, ta có thể sử dụng thông tin về phân phối biên hoặc thông tin phụ thuộc chung.

Chương 3

Ứng dụng

Để hiểu rõ hơn về ứng dụng của các phương pháp hiệu chỉnh đã trình bày ở trên cho bài toán kiểm định bội trong thực tế, phần sau đây chúng ta thực hành kiểm định đa giả thuyết với bộ dữ liệu thực tế về bệnh bạch cầu. Kiểm định bộ dữ liệu được thực hiện với ngôn ngữ lập trình R.

Trong phần này em sẽ chạy dữ liệu với 3 phương pháp hiệu chỉnh *p-giá trị* là các phương pháp Bonferroni, Holm và Benjamini - Hochberg. Về *e*-BH, là một trường hợp tổng quát hơn của phương pháp Benjamini - Hochberg, trong báo cáo này chỉ mang tính giới thiệu và chưa chạy số liệu do chưa thể trình bày được cụ thể về *e-giá trị*.

3.1 Giới thiệu bộ dữ liệu

Bộ dữ liệu được lấy từ nghiên cứu của Golub et al. (1999) về 2 chủng bệnh bạch cầu. Dữ liệu thể hiện mức độ biểu hiện của một số gen xác định ở các bệnh nhân bị bệnh bạch cầu. Trong nghiên cứu này, các tác giả quan tâm đến sự khác nhau về mức độ biểu hiện của các gen đó ở trên 2 chủng bệnh là bệnh bạch cầu ác tính dòng lympho (ALL) và bệnh bạch cầu ác tính dòng myeloid (AML). Cụ thể, bộ dữ liệu ghi lại mức độ biểu hiện của 3051 gen khác nhau trên 38 mẫu, là

38 khối u ác tính ở các bệnh nhân thuộc 2 chủng bệnh, trong đó có 27 trường hợp ALL và 11 trường hợp AML. Dữ liệu đã được thực hiện qua một số bước tiền xử lý để thu được ma trận chuẩn hóa về giá trị mức độ biểu hiện mà ta sử dụng. Chi tiết thông tin trong bộ dữ liệu như sau:

- `golub`: ma trận 3051×38 ghi lại mức độ biểu hiện của gen trên 38 mẫu thu được

```
> golub[1:5, 1:5]
          [,1]     [,2]     [,3]     [,4]     [,5]
[1,] -1.45769 -1.39420 -1.42779 -1.40715 -1.42668
[2,] -0.75161 -1.26278 -0.09052 -0.99596 -1.24245
[3,]  0.45695 -0.09654  0.90325 -0.07194  0.03232
[4,]  3.13533  0.21415  2.08754  2.23467  0.93811
[5,]  2.76569 -1.27045  1.60433  1.53182  1.63728
```

Hình 3.1: Dữ liệu `golub`

- `golub.gnames`: ma trận 3051×3 chứa thông tin xác định của gen, gồm số thứ tự, mã số khoa học và tên

```
> golub.gnames[1:100, ]
      [,1]  [,2]
  [1,] "36"  "AFFX-HUMISGF3A/M97935_MA_at (endogenous control)"
  [2,] "37"  "AFFX-HUMISGF3A/M97935_MB_at (endogenous control)"
  [3,] "38"  "AFFX-HUMISGF3A/M97935_3_at (endogenous control)"
  [4,] "39"  "AFFX-HUMRGE/M10098_5_at (endogenous control)"
  [5,] "40"  "AFFX-HUMRGE/M10098_M_at (endogenous control)"
  [6,] "41"  "AFFX-HUMRGE/M10098_3_at (endogenous control)"
  [7,] "42"  "AFFX-HUMGAPDH/M33197_5_at (endogenous control)"
  [8,] "43"  "AFFX-HUMGAPDH/M33197_M_at (endogenous control)"
  [9,] "45"  "AFFX-HSAC07/X00351_5_at (endogenous control)"
 [10,] "46"  "AFFX-HSAC07/X00351_M_at (endogenous control)"
 [11,] "48"  "AFFX-HUMTFRR/M11507_5_at (endogenous control)"
 [12,] "49"  "AFFX-HUMTFRR/M11507_M_at (endogenous control)"
 [13,] "50"  "AFFX-HUMTFRR/M11507_3_at (endogenous control)"
 [14,] "51"  "AFFX-M27830_5_at (endogenous control)"
 [15,] "52"  "AFFX-M27830_M_at (endogenous control)"
 [16,] "53"  "AFFX-M27830_3_at (endogenous control)"
 [17,] "54"  "AFFX-HSAC07/X00351_3_st (endogenous control)"
 [18,] "56"  "AFFX-HUMGAPDH/M33197_M_st (endogenous control)"
 [19,] "57"  "AFFX-HUMGAPDH/M33197_3_st (endogenous control)"
 [20,] "59"  "AFFX-HSAC07/X00351_M_st (endogenous control)"
 [21,] "60"  "GB DEF = GABAa receptor alpha-3 subunit"
 [22,] "62"  "mRNA"
 [23,] "65"  "VRK1"
 [24,] "69"  "mRNA, clone RES4-24A, exon 1, 2, 3, 4"
 [25,] "72"  "Zinc finger protein, clone RES4-26"
 [26,] "73"  "Prostate differentiation factor mRNA"
 [27,] "78"  "GLIA MATURATION FACTOR BETA"
 [28,] "79"  "AQP3 Aquaporin 3"
 [29,] "81"  "KIAA0317 gene"
 [30,] "82"  "KIAA0320 gene, partial cds"
 [31,] "86"  "KIAA0384 gene"
 [32,] "88"  "Hunc18b2"
 [33,] "91"  "Proteasome subunit p27"
 [34,] "95"  "WUGSC:H_133K23.1c gene extracted from Human BAC clone 133K23 from 7q31.2"
 [35,] "98"  "WUGSC:H_RG083M05.2 gene extracted from Human BAC clone RG083M05 from 7q21-7q22"
 [36,] "104" "WUGSC:DJ515N1.2 gene extracted from Human PAC clone DJ515N1 from 22q11.2-q22"
```

Hình 3.2: Dữ liệu `golub.gnames`

- `golub.cl`: vector chứa nhãn phân loại bệnh nhân (0 là trường hợp ALL, 1 là trường hợp AML)

```
> golub.cl
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
```

Hình 3.3: Dữ liệu `golub.cl`

Bài toán đặt ra là, trong số các gen được nghiên cứu, những gen nào có mức độ biểu hiện là như nhau ở 2 chủng bệnh, và những gen nào có mức độ biểu hiện khác nhau? Hay nói cách khác, ta cần phải kiểm định 3051 giả thuyết H_0:

$$H_0 : \mu_1 = \mu_2,$$

với μ_1 là mức độ biểu hiện trung bình của một gen ở những người bệnh ALL,

còn μ_2 là mức độ biểu hiện trung bình của gen đó ở người bệnh AML.

3.2 Các bước thực hiện

Trước tiên, ta tính thống kê t so sánh 2 mẫu của từng giả thuyết với công thức:

$$t = \frac{\mu_1 - \mu_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ là sai số hiệu chỉnh gộp của 2 nhóm bệnh, n_1, n_2 lần lượt là số quan sát của trường hợp ALL và AML, s_1, s_2 lần lượt là sai số hiệu chỉnh của các trường hợp ALL và AML. Thống kê t_i ứng với giả thuyết thứ i là H_i.

Thực hiện chạy dữ liệu ta tính được thống kê t của 3051 giả thuyết. Thống kê t của 100 gen đầu tiên tính được như sau:

```
> teststat[1:100]
  [1]  1.76  0.91 -0.10 -0.34 -1.37 -1.27  0.72  0.53  0.02  0.13
 [11]  4.13  2.85  3.92 -0.44 -0.19  0.35 -1.86  2.61  0.40  2.22
 [21]  1.45  0.45 -4.51 -0.69 -2.20  1.43 -0.38 -0.05  1.29 -0.44
 [31] -0.13  3.66 -0.27 -1.14  2.72  2.52 -1.08  0.39 -3.19  1.61
 [41]  1.64  1.12  2.62 -0.98  0.90  1.05 -1.47 -1.72  1.16  3.20
 [51] -2.69 -0.59  1.50 -0.81 -4.07  4.29  0.64 -0.02 -1.87  2.45
 [61] -1.65  3.29 -1.87  1.13 -0.88 -4.67  1.88  5.18 -0.95  0.60
 [71] -0.54  1.26  1.61 -2.79 -1.24  0.27 -1.98 -2.86 -2.45 -1.19
 [81] -3.42  2.25 -1.49 -3.61  0.58  0.07 -0.01 -0.89 -2.06  1.36
 [91]  0.12 -0.71 -0.76 -0.84 -0.51 -5.87  0.71 -1.08  0.13  1.78
```

Hình 3.4: Thống kê t của 100 gen đầu tiên

Ước lượng *p-giá trị* ban đầu của 3051 giả thuyết cho các thống kê t đã tính được, sử dụng phân phối chuẩn.

P-giá trị ban đầu của 100 giả thuyết đầu tiên ước lượng được như sau:

```
> rawp[1:100]
  [1] 0.08 0.36 0.92 0.73 0.17 0.20 0.47 0.60 0.98 0.90
 [11] 0.00 0.00 0.00 0.66 0.85 0.73 0.06 0.01 0.69 0.03
 [21] 0.15 0.65 0.00 0.49 0.03 0.15 0.70 0.96 0.20 0.66
 [31] 0.90 0.00 0.79 0.25 0.01 0.01 0.28 0.70 0.00 0.11
 [41] 0.10 0.26 0.01 0.33 0.37 0.29 0.14 0.09 0.25 0.00
 [51] 0.01 0.56 0.13 0.42 0.00 0.00 0.52 0.98 0.06 0.01
 [61] 0.10 0.00 0.06 0.26 0.38 0.00 0.06 0.00 0.34 0.55
 [71] 0.59 0.21 0.11 0.01 0.21 0.79 0.05 0.00 0.01 0.23
 [81] 0.00 0.02 0.14 0.00 0.56 0.94 0.99 0.37 0.04 0.17
 [91] 0.90 0.48 0.45 0.40 0.61 0.00 0.48 0.28 0.90 0.08
```

Hình 3.5: *P-giá trị* ban đầu của 100 giả thuyết đầu tiên

Hiệu chỉnh *p-giá trị* ban đầu theo 3 thuật toán đã trình bày, thu được các *p-giá trị* hiệu chỉnh theo từng phương pháp như sau:

	rawp	adjp_B	adjp_H	adjp_BH
1	0.08	1	1	0.19
2	0.36	1	1	0.53
3	0.92	1	1	0.96
4	0.73	1	1	0.84
5	0.17	1	1	0.32
6	0.20	1	1	0.36
7	0.47	1	1	0.64
8	0.60	1	1	0.75
9	0.98	1	1	0.99
10	0.90	1	1	0.95
11	0.00	0	0	0.00
12	0.00	0	0	0.00
13	0.00	0	0	0.00
14	0.66	1	1	0.79
15	0.85	1	1	0.92
16	0.73	1	1	0.84
17	0.06	1	1	0.15
18	0.01	1	1	0.04
19	0.69	1	1	0.81
20	0.03	1	1	0.09
21	0.15	1	1	0.30
22	0.65	1	1	0.78
23	0.00	0	0	0.00
24	0.49	1	1	0.66
25	0.03	1	1	0.09
26	0.15	1	1	0.30
27	0.70	1	1	0.82
28	0.96	1	1	0.98
29	0.20	1	1	0.36
30	0.66	1	1	0.79
31	0.90	1	1	0.95
32	0.00	0	0	0.00
33	0.79	1	1	0.88
34	0.25	1	1	0.42
35	0.01	1	1	0.04

Hình 3.6: *P-giá trị* ban đầu

	reject_raw	reject_B	reject_H	reject_BH
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	1	0	0	1
19	0	0	0	0
20	1	0	0	0
21	0	0	0	0
22	0	0	0	0
23	1	1	1	1
24	0	0	0	0
25	1	0	0	0
26	0	0	0	0
27	0	0	0	0
28	0	0	0	0
29	0	0	0	0
30	0	0	0	0
31	0	0	0	0
32	1	1	1	1
33	0	0	0	0
34	0	0	0	0
35	1	0	0	1
36	1	0	0	1

Hình 3.7: Ma trận bác bỏ

Số giả thuyết bị bác bỏ theo mỗi phương pháp là:

	Bonferroni	Holm	Benjamini - Hochberg
Số giả thuyết bị bác bỏ	686	686	894

Những gen có mức độ biểu hiện khác nhau ở 2 chủng bệnh:

```
> genes_different_B = golub.gnames[genes_different_B, ]
> genes_different_B[1:100, ]
      [,1]  [,2]                                                                    [,3]
 [1,] "48"  "AFFX-HUMTFRR/M11507_5_at (endogenous control)"                        "AFFX-HUMTFRR/M11507_5_at"
 [2,] "49"  "AFFX-HUMTFRR/M11507_M_at (endogenous control)"                        "AFFX-HUMTFRR/M11507_M_at"
 [3,] "50"  "AFFX-HUMTFRR/M11507_3_at (endogenous control)"                        "AFFX-HUMTFRR/M11507_3_at"
 [4,] "65"  "VRK1"                                                                 "AB000449_at"
 [5,] "88"  "Hunc18b2"                                                             "AB002559_at"
 [6,] "119" "RAS-RELATED PROTEIN RAB-11A"                                          "AF000231_at"
 [7,] "146" "GB DEF = Syntaxin-16C mRNA"                                           "AF008937_at"
 [8,] "155" "GB DEF = SKB1Hs mRNA"                                                 "AF015913_at"
 [9,] "157" "GB DEF = C8FW phosphoprotein"                                         "AJ000480_at"
[10,] "167" "FECH Ferrochelatase (protoporphyria)"                                 "D00726_at"
[11,] "171" "GAPD Glyceraldehyde-3-phosphate dehydrogenase"                        "D00763_at"
[12,] "173" "PRKCD Protein kinase C, delta"                                        "D10495_at"
[13,] "196" "DNA-(APURINIC OR APYRIMIDINIC SITE) LYASE"                            "D13370_at"
[14,] "200" "KIAA0002 gene"                                                        "D13627_at"
[15,] "206" "KIAA0011 gene"                                                        "D13636_at"
[16,] "229" "KIAA0102 gene"                                                        "D14658_at"
[17,] "242" "KIAA0110 gene"                                                        "D14811_at"
[18,] "248" "ADM Adrenomedullin"                                                   "D14874_at"
[19,] "259" "ORF, Xq terminal portion"                                             "D16469_at"
[20,] "275" "KIAA0030 gene, partial cds"                                           "D21063_at"
[21,] "283" "SM22-ALPHA HOMOLOG"                                                   "D21261_at"
[22,] "284" "KIAA0035 gene, partial cds"                                           "D21262_at"
[23,] "288" "KIAA0029 gene, partial cds"                                           "D21852_at"
[24,] "302" "PFKP Phosphofructokinase, platelet"                                   "D25328_at"
[25,] "309" "KIAA0041 gene, partial cds"                                           "D26069_at"
[26,] "312" "NADPH-flavin reductase"                                               "D26308_at"
[27,] "321" "Transmembrane protein"                                                "D26579_at"
[28,] "341" "HYPOTHETICAL MYELOID CELL LINE PROTEIN 2"                             "D29641_at"
[29,] "343" "KIAA0115 gene"                                                        "D29643_at"
[30,] "353" "EIF4A2 Eukaryotic translation initiation factor 4A (eIF-4A) isoform 2" "D30655_at"
[31,] "361" "KIAA0064 gene"                                                        "D31764_at"
[32,] "370" "KIAA0063 gene"                                                        "D31884_at"
[33,] "376" "KIAA0067 gene"                                                        "D31891_at"
[34,] "379" "AARS Alanyl-tRNA synthetase"                                          "D32050_at"
[35,] "385" "Proteasome subunit z"                                                 "D38048_at"
[36,] "396" "KIAA0077 gene, partial cds"                                           "D38521_at"
```

Hình 3.8: Các gen có biểu hiện khác nhau theo Bonferroni

```
> genes_different_H = golub.gnames[genes_different_H, ]
> genes_different_H[1:100, ]
      [,1]  [,2]                                                                  [,3]
 [1,] "48"  "AFFX-HUMTFRR/M11507_5_at (endogenous control)"                       "AFFX-HUMTFRR/M11507_5_at"
 [2,] "49"  "AFFX-HUMTFRR/M11507_M_at (endogenous control)"                      "AFFX-HUMTFRR/M11507_M_at"
 [3,] "50"  "AFFX-HUMTFRR/M11507_3_at (endogenous control)"                      "AFFX-HUMTFRR/M11507_3_at"
 [4,] "65"  "VRK1"                                                               "AB000449_at"
 [5,] "88"  "Hunc18b2"                                                           "AB002559_at"
 [6,] "119" "RAS-RELATED PROTEIN RAB-11A"                                        "AF000231_at"
 [7,] "146" "GB DEF = Syntaxin-16C mRNA"                                         "AF008937_at"
 [8,] "155" "GB DEF = SKB1Hs mRNA"                                               "AF015913_at"
 [9,] "157" "GB DEF = C8FW phosphoprotein"                                       "AJ000480_at"
[10,] "167" "FECH Ferrochelatase (protoporphyria)"                               "D00726_at"
[11,] "171" "GAPD Glyceraldehyde-3-phosphate dehydrogenase"                      "D00763_at"
[12,] "173" "PRKCD Protein kinase C, delta"                                      "D10495_at"
[13,] "196" "DNA-(APURINIC OR APYRIMIDINIC SITE) LYASE"                          "D13370_at"
[14,] "200" "KIAA0002 gene"                                                      "D13627_at"
[15,] "206" "KIAA0011 gene"                                                      "D13636_at"
[16,] "229" "KIAA0102 gene"                                                      "D14658_at"
[17,] "242" "KIAA0110 gene"                                                      "D14811_at"
[18,] "248" "ADM Adrenomedullin"                                                 "D14874_at"
[19,] "259" "ORF, Xq terminal portion"                                           "D16469_at"
[20,] "275" "KIAA0030 gene, partial cds"                                         "D21063_at"
[21,] "283" "SM22-ALPHA HOMOLOG"                                                 "D21261_at"
[22,] "284" "KIAA0035 gene, partial cds"                                         "D21262_at"
[23,] "288" "KIAA0029 gene, partial cds"                                         "D21852_at"
[24,] "302" "PFKP Phosphofructokinase, platelet"                                 "D25328_at"
[25,] "309" "KIAA0041 gene, partial cds"                                         "D26069_at"
[26,] "312" "NADPH-flavin reductase"                                             "D26308_at"
[27,] "321" "Transmembrane protein"                                              "D26579_at"
[28,] "341" "HYPOTHETICAL MYELOID CELL LINE PROTEIN 2"                           "D29641_at"
[29,] "343" "KIAA0115 gene"                                                      "D29643_at"
[30,] "353" "EIF4A2 Eukaryotic translation initiation factor 4A (eIF-4A) isoform 2" "D30655_at"
[31,] "361" "KIAA0064 gene"                                                      "D31764_at"
[32,] "370" "KIAA0063 gene"                                                      "D31884_at"
[33,] "376" "KIAA0067 gene"                                                      "D31891_at"
[34,] "379" "AARS Alanyl-tRNA synthetase"                                        "D32050_at"
[35,] "385" "Proteasome subunit z"                                               "D38048_at"
[36,] "396" "KIAA0077 gene, partial cds"                                         "D38521_at"
```

Hình 3.9: Các gen có biểu hiện khác nhau theo Holm

```
> genes_different_BH = golub.gnames[genes_different_BH, ]
> genes_different_BH[1:100, ]
      [,1]  [,2]                                                                  [,3]
 [1,] "48"  "AFFX-HUMTFRR/M11507_5_at (endogenous control)"                       "AFFX-HUMTFRR/M11507_5_at"
 [2,] "49"  "AFFX-HUMTFRR/M11507_M_at (endogenous control)"                      "AFFX-HUMTFRR/M11507_M_at"
 [3,] "50"  "AFFX-HUMTFRR/M11507_3_at (endogenous control)"                      "AFFX-HUMTFRR/M11507_3_at"
 [4,] "56"  "AFFX-HUMGAPDH/M33197_M_st (endogenous control)"                     "AFFX-HUMGAPDH/M33197_M_st"
 [5,] "65"  "VRK1"                                                               "AB000449_at"
 [6,] "88"  "Hunc18b2"                                                           "AB002559_at"
 [7,] "98"  "WUGSC:H_RG083M05.2 gene extracted from Human BAC clone RG083M05 from 7q21-7q22"  "AC000064_cds1_at"
 [8,] "104" "WUGSC:DJ515N1.2 gene extracted from Human PAC clone DJ515N1 from 22q11.2-q22"    "AC002073_cds1_at"
 [9,] "119" "RAS-RELATED PROTEIN RAB-11A"                                        "AF000231_at"
[10,] "131" "RET ligand 2 (RETL2) mRNA"                                          "AF002700_at"
[11,] "146" "GB DEF = Syntaxin-16C mRNA"                                         "AF008937_at"
[12,] "147" "GB DEF = TEB4 protein mRNA"                                         "AF009301_at"
[13,] "155" "GB DEF = SKB1Hs mRNA"                                               "AF015913_at"
[14,] "157" "GB DEF = C8FW phosphoprotein"                                       "AJ000480_at"
[15,] "164" "GPX3 Glutathione peroxidase 3 (plasma)"                             "D00632_at"
[16,] "167" "FECH Ferrochelatase (protoporphyria)"                               "D00726_at"
[17,] "171" "GAPD Glyceraldehyde-3-phosphate dehydrogenase"                      "D00763_at"
[18,] "173" "PRKCD Protein kinase C, delta"                                      "D10495_at"
[19,] "181" "IL2RG Interleukin 2 receptor gamma chain"                           "D11086_at"
[20,] "196" "DNA-(APURINIC OR APYRIMIDINIC SITE) LYASE"                          "D13370_at"
[21,] "197" "PIGF Phosphatidylinositol glycan, class F"                          "D13435_at"
[22,] "200" "KIAA0002 gene"                                                      "D13627_at"
[23,] "206" "KIAA0011 gene"                                                      "D13636_at"
[24,] "229" "KIAA0102 gene"                                                      "D14658_at"
[25,] "239" "PHOSPHATIDYLSERINE SYNTHASE I"                                      "D14694_at"
[26,] "241" "ATP5A1 ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle" "D14710_at"
[27,] "242" "KIAA0110 gene"                                                      "D14811_at"
[28,] "248" "ADM Adrenomedullin"                                                 "D14874_at"
[29,] "250" "Small GTP-binding protein, S10"                                     "D14889_at"
[30,] "253" "DEFENDER AGAINST CELL DEATH 1"                                      "D15057_at"
[31,] "255" "CAST Calpastatin"                                                   "D16217_at"
[32,] "259" "ORF, Xq terminal portion"                                           "D16469_at"
[33,] "261" "ATP SYNTHASE GAMMA CHAIN, MITOCHONDRIAL PRECURSOR"                  "D16562_at"
[34,] "275" "KIAA0030 gene, partial cds"                                         "D21063_at"
[35,] "283" "SM22-ALPHA HOMOLOG"                                                 "D21261_at"
[36,] "284" "KIAA0035 gene, partial cds"                                         "D21262_at"
```

Hình 3.10: Các gen có biểu hiện khác nhau theo Benjamini - Hochberg

3.3 Nhận xét

Với bộ dữ liệu về mức độ biểu hiện của 3051 gen quan sát được trên 38 mẫu, được chia thành 2 nhóm ALL và AML, sau khi thực hiện theo 3 quy trình để kiểm định giả thuyết đã nêu trên, với mức ý nghĩa $\alpha = 0.05$, ta thu được kết quả:

- Thủ tục Bonferroni và Holm tìm được số lượng gen có biểu hiện khác nhau ở 2 nhóm ALL và AML là như nhau, 686 gen.
- Phương pháp Benjamini - Hochberg đưa ra kết quả là 894 gen có biểu hiện khác nhau ở 2 nhóm ALL và AML, nhiều hơn so với 2 phương pháp kiểm soát $FWER$.

Từ đó ta có được một số nhận xét:

- Cả 3 phương pháp giúp kiểm soát được tỷ lệ sai lầm loại 1, tức là giảm khả năng kết luận nhầm một gen có biểu hiện giống nhau ở 2 nhóm thành gen có biểu hiện khác thường.
- Phương pháp Benjamini - Hochberg đưa ra kết quả có số lượng lớn hơn 2 phương pháp còn lại. Trong số hơn 200 gen nhiều hơn này, có thể có một vài kết quả sai, tức là một số trường hợp được kết luận là gen có biểu hiện như nhau ở 2 nhóm, thực ra là có biểu hiện khác nhau, tuy nhiên số lượng này thấp ở mức chấp nhận được. Bù lại, lực lượng của kiểm định được tăng lên, tức là ta tìm thêm được một số gen có biểu hiện khác biệt ở 2 nhóm mà các phương pháp kiểm soát $FWER$ đã bỏ qua.

Chương 4

Kết luận

Các kết quả đạt được của đồ án

Đồ án đã đạt được các mục tiêu đề ra, gồm các nội dung được tóm tắt như sau:

1. Trình bày được những khái niệm cơ bản của bài toán kiểm định bội.
2. Trình bày được một số phương pháp hiệu chỉnh tiêu biểu: phương pháp Bonferroni và phương pháp Holm kiểm soát $FWER$, phương pháp Benjamini - Hochberg và e-BH kiểm soát FDR.
3. Ứng dụng được các thuật toán trên vào dữ liệu thực tế, thực hiện qua chương trình chạy bằng ngôn ngữ R.

Hướng phát triển của đồ án trong tương lai

1. Tìm hiểu sâu hơn về *e-giá trị*, chi tiết thuật toán e-BH và các thủ tục hiệu chỉnh *e-giá trị*.
2. Tìm hiểu và xây dựng để phát triển một phương pháp mới tối ưu hơn, vừa có thể kiểm soát số sai lầm loại 1 chặt chẽ hơn, vừa làm tăng lực lượng kiểm định ở mức tốt hơn dựa trên các phương pháp đã được đưa ra.

Tài liệu tham khảo

[1.] Golub et al, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 286, 1999, 531-537.

[2.] Ruodu Wang, "Elementary proofs of several results on false discovery rate", 2022.

[3.] Ruodu Wang, Aaditya Ramdas, "False discovery rate control with e-values", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3), 2022, 822–852.

[4.] Sandrine Dudoit, Yongchao Ge, "Bioconductor's multtest package", 2003.

[5.] Sture Holm, "A Simple Sequentially Rejective Multiple Test Procedure", *Scandinavian Journal of Statistic*, 6(2), 1979, 65-70.

[6.] Yoav Benjamini, Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1), 1995, 289-300.

PHỤ LỤC

Mã nguồn chương trình bằng ngôn ngữ R

```
# nhap thu vien va du lieu
library(multtest, verbose = FALSE)
library(dplyr)
data(golub)
# quan sat du lieu
golub[1:5, 1:5]
golub.gnames[1:5, ]
golub.cl
# tinh thong ke t
teststat = mt.teststat(golub, golub.cl)
teststat = round(teststat, 2)
teststat[1:100]
# tinh p-value ban dau
rawp = 2*(1-pnorm(abs(teststat)))
rawp = round(rawp, 2)
rawp[1:100]
df_rawp = data.frame(rawp)
# ham hieu chinh Bonferroni
Bonferroni = function(vector){
```

```
  m = length(vector)
  for (i in 1:m){
    vector[i] = min(vector[i]*m, 1)
  }
  return(vector)
}
adjp_B = Bonferroni(rawp)
adjp_B = round(adjp_B, 2)
df_adjp_B = data.frame(adjp_B)
# ham hieu chinh Holm
Holm = function(vector){
  m = length(vector)
  sorted = sort(vector, decreasing = FALSE)
  match = match(vector, sorted)
  for (i in 1:m){
    sorted[i] = min(sorted[i]*(m+1-i), 1)
  }
  sorted = sorted[match]
  return(sorted)
}
adjp_H = Holm(rawp)
adjp_H = round(adjp_H, 2)
df_adjp_H = data.frame(adjp_H)
# ham hieu chinh Benjamini_Hochberg
Benjamini_Hochberg = function(vector){
  m = length(vector)
  sorted = sort(vector, decreasing = FALSE)
```

```
47    match = match(vector, sorted)
48    for (i in 1:m){
49      sorted[i] = sorted[i]*(m/i)
50    }
51    sorted = sorted[match]
52    return(sorted)
53  }
54  adjp_BH = Benjamini_Hochberg(rawp)
55  adjp_BH = round(adjp_BH, 2)
56  df_adjp_BH = data.frame(adjp_BH)
57  # in ra p-value hieu chinh
58  summarize = bind_cols(df_rawp,df_adjp_B,df_adjp_H,df_
      adjp_BH)
59  summarize[1:100, ]
60  # ham bac bo
61  reject = function(vector, alpha){
62    m = length(vector)
63    for (i in m:1){
64      if (vector[i]>alpha) {
65        vector[i] = 0
66      }
67      else {
68        vector[i] = 1
69      }
70    }
71    return(vector)
72  }
```

```
reject_raw = reject(rawp,0.05)
reject_B = reject(adjp_B,0.05)
reject_H = reject(adjp_H,0.05)
reject_BH = reject(adjp_BH,0.05)
df_reject_raw = data.frame(reject_raw)
df_reject_B = data.frame(reject_B)
df_reject_H = data.frame(reject_H)
df_reject_BH = data.frame(reject_BH)
rejected = bind_cols(df_reject_raw, df_reject_B, df_
   reject_H, df_reject_BH)
rejected[1:100, ]
# in ra gen co bieu hien khac
no_reject_B = 0
genes_different_B = c()
for (i in 1:m){
  if (reject_B[i] == 1) {
    no_reject_B = no_reject_B + 1
    genes_different_B[no_reject_B] = i
  }
  else next
}
genes_different_B
no_reject_B

no_reject_H = 0
genes_different_H = c()
for (i in 1:m){
```

```
  if (reject_H[i] == 1) {
    no_reject_H = no_reject_H + 1
    genes_different_H[no_reject_H] = i
  }
  else next
}
genes_different_H
no_reject_H

no_reject_BH = 0
genes_different_BH = c()
for (i in 1:m){
  if (reject_BH[i] == 1) {
    no_reject_BH = no_reject_BH + 1
    genes_different_BH[no_reject_BH] = i
  }
  else next
}
genes_different_BH
no_reject_BH
```