Bài giảng 4: A/B Testing

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Học phần: Xử lý Số liệu Thống kê

Nội dung buổi học

1 Kiểm định giả thuyết

2 A/B testing hai nhóm

3 A/B testing nhiều nhóm

4 Multi-Armed Bandit

1 Kiểm định giả thuyết

2 A/B testing hai nhóm

3 A/B testing nhiều nhóm

4 Multi-Armed Bandit

Khái niệm cơ bản

Kiểm định giả thuyết là một bài toán cơ bản trong thống kê lý thuyết cũng như thống kê ứng dụng.

Thông thường, kiểm định giả thuyết hướng tới trả lời một giả thuyết cho 1 đặc trưng của dữ liệu:

- Trung bình đối với biến định lượng;
- Tỷ lệ đối với biến định tính;
- Sự tương quan giữa hai biến.

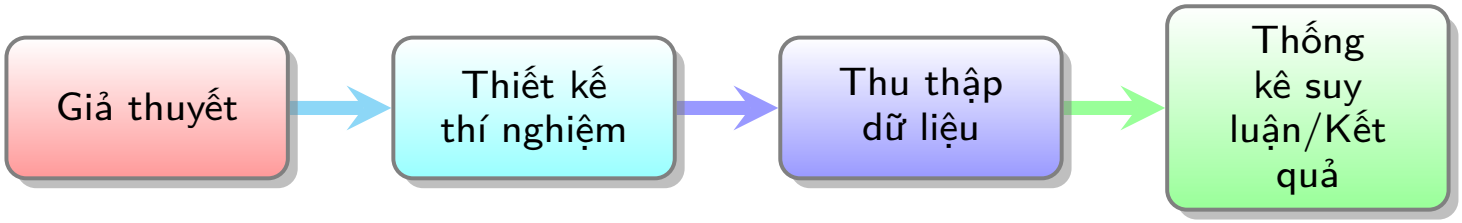
Khái niệm cơ bản

Kiểm định giả thuyết là một bài toán cơ bản trong thống kê lý thuyết cũng như thống kê ứng dụng.

Thông thường, kiểm định giả thuyết hướng tới trả lời một giả thuyết cho 1 đặc trưng của dữ liệu:

- Trung bình đối với biến định lượng;
- Tỷ lệ đối với biến định tính;
- Sự tương quan giữa hai biến.

Quy trình kiểm định giả thuyết thống kê:



Khái niệm cơ bản

Một giả thuyết thống kê gồm 2 thành phần:

Giả thuyết - Null hypothesis là một giả định bao hàm tính không khác biệt/không hiệu quả hoặc sự khác biệt chỉ là sự trùng hợp ngẫu nhiên.

Ví dụ: "Mức độ tương tác của trang web A và trang web B là không khác biệt."

Đối thuyết - Alternative hypothesis là một giả định bao hàm sự ngược lại giả định đã được tuyên bố.

Ví dụ: "Mức độ tương tác của trang web A và trang web B là khác biệt."

"Mức độ tương tác của trang web A là cao hơn của trang web B."

"Mức độ tương tác của trang web A là thấp hơn của trang web B."

Thông thường, ta mong muốn chứng minh rằng giả thuyết là **không đúng**, thông qua dữ liệu thu thập từ việc thiết kế thí nghiệm.

A/B testing

Một kiểm định A/B (A/B testing) là một thí nghiệm với hai nhóm nhằm xác định xem nhóm nào trong: hai phương pháp điều trị, hay hai sản phẩm, hay hai quy trình là tốt hơn cái còn lại.

Thông thường, hai nhóm gồm:

nhóm đối chứng - control group là nhóm gồm các đối tượng được tiếp xúc với điều kiện thông thường (không được áp dụng các biện pháp mới) - được ký hiệu là A;

nhóm điều trị - treatment group là nhóm gồm các đối tượng được tiếp xúc với 1 điều kiện điều trị cụ thể - được ký hiệu là B.

Giả thuyết và đối thuyết của A/B testing thường là

- Giả thuyết: Phương pháp điều trị có hiệu quả tương đồng với đối chứng.
- Đối thuyết: Phương pháp điều trị có hiệu quả tốt hơn đối chứng.

A/B testing

Ví dụ cho A/B testing:

- Thử nghiệm hai phương pháp xử lý đất để xác định phương pháp nào giúp hạt nảy mầm tốt hơn.
- Thử nghiệm hai liệu pháp để xác định liệu pháp nào ngăn chặn ung thư hiệu quả hơn.
- Kiểm tra hai mức giá để xác định mức giá nào mang lại nhiều lợi nhuận ròng hơn.
- Kiểm tra hai dòng tiêu đề trên web để xác định dòng tiêu đề nào tạo ra nhiều lượt nhấp chuột hơn.
- Thử nghiệm hai quảng cáo trên web để xác định quảng cáo nào tạo ra nhiều chuyển đổi hơn.

Đối tượng nghiên cứu trong một A/B test, có thể là:

- người;
- hạt mầm cây;
- người ghé thăm trang web.

A/B testing

Mỗi đối tượng sẽ được tiếp xúc với

- điều kiện thông thường (control); hoặc
- điều kiện điều trị (treatment).

Để có dữ liệu phục vụ cho việc phân tích, ta cần thiết kế thí nghiệm đảm bảo tính ngẫu nhiên (randomness).

$\hookrightarrow$ Các đối tượng được chỉ định một cách ngẫu nhiên (**randomization**) vào nhóm đối chứng hoặc nhóm điều trị.

$\hookrightarrow$ Khi đó, ta sẽ biết rằng bất kỳ sự khác biệt nào giữa các nhóm là do một trong hai điều sau:

- Hiệu quả của các phương pháp điều trị khác nhau.
- Sự may mắn trong kết quả do các đối tượng được chỉ định ngẫu nhiên vào các phương pháp điều trị (tức là, việc phân công ngẫu nhiên có thể dẫn đến việc các đối tượng có thành tích tốt hơn một cách tự nhiên tập trung vào A hoặc B).

Mở rộng A/B testing

A/B testing là khá phổ biến trong các lĩnh vực kinh doanh hoặc thương mại điện tử.

Tuy nhiên, vẫn có những loại kiểm định khác, là biến thể hoặc mở rộng của A/B testing:

- Số nhóm điều trị nhiều hơn hai.

  Ví dụ: so sánh độ gắn kết khách hàng của 4 trang web.
- Sự liên hệ của hai biến định tính.

  Ví dụ: kiểm tra sự tác động của 3 kiểu headline tới tương tác của người dùng (click hoặc no-click).
- Các đối tượng được quan sát lặp lại.

Mở rộng A/B testing

Trong thống kê, ta thường quan tâm tới các câu hỏi dạng:

"Sự khác biệt giữa phương pháp điều trị A và phương pháp điều trị B có ý nghĩa thống kê hay không?"

Tuy nhiên, các nhà khoa học dữ liệu quan tâm tới câu hỏi dạng:

"Trong số nhiều phương pháp điều trị khác nhau, cái nào là tốt nhất?"

tức là dạng câu hỏi trực diện hơn.

Với câu hỏi dạng này, một loại thiết kế thử nghiệm tương đối mới được sử dụng: the multi-armed bandit hay **Multi-Armed Bandit Algorithms**.

Sai lầm trong kiểm định

Bảng sau đây biểu diễn bốn hậu quả có thể của một quyết định trong bài toán kiểm định thống kê:

Quyết định	Sự thật của Giả thuyết: Đúng	Sai
Không bác bỏ	Chính xác	Sai lầm loại II
Bác bỏ	Sai lầm loại I	Chính xác

Sai lầm loại I - Type I error

Sai lầm loại I là sai lầm có thể mắc phải trong kiểm định giả thuyết thống kê, nó xảy ra khi ta quyết định **bác bỏ** Giả thuyết khi Giả thuyết thực sự đúng.

Sai lầm loại II - Type II error

Sai lầm loại II là sai lầm có thể mắc phải trong kiểm định giả thuyết thống kê, nó xảy ra khi ta quyết định **không bác bỏ** Giả thuyết khi Giả thuyết thực sự sai.

Sai lầm trong kiểm định

- Xác suất mắc sai lầm loại I thường được ký hiệu là α và nó là **mức ý nghĩa**.
- Xác suất mắc sai lầm loại II thường được ký hiệu là β.
- $1 - \beta$ được gọi là độ mạnh của kiểm định (power), tức là khả năng nhận ra được Giả thuyết là sai.

Trong khuôn khổ một kiểm định thống kê, chúng ta giả định rằng Giả thuyết là đúng. Vì vậy, chúng ta có thể nhận được:

- quyết định chính xác nếu ta không bác bỏ Giả thuyết; hoặc,
- sai lầm loại I nếu ta bác bỏ Giả thuyết.

Để kiểm soát khả năng bị sai lầm loại I, ta cố định xác suất mắc sai lầm loại I (*mức ý nghĩa*), thường là 0.05 (hay 5%) và so sánh nó với một đại lượng được gọi là *p*-value, được tính từ dữ liệu.

Trong khi đó, sai lầm loại II được kiểm soát bằng việc xác định cỡ mẫu của dữ liệu.

p-value và quyết định

p-value là xác suất mà mô hình kiểm định tạo ra kết quả cực đoan hơn kết quả quan sát được, dưới giả định rằng Giả thuyết là đúng.

Ví dụ: *p*-value bằng 0.308 có nghĩa là ta có thể kỳ vọng rằng, ta sẽ thu được kết quả cực đoan như của dữ liệu được quan sát hoặc cực đoan hơn, một cách ngẫu nhiên trong hơn 30% số lần lấy mẫu, khi giả sử rằng Giả thuyết là đúng.

Nói cách khác, *p*-value có thể được dùng để do sự tương thích của dữ liệu với Giả thuyết.

- nếu *p*-value là lớn, có nghĩa là dữ liệu tương thích với Giả thuyết
  $\hookrightarrow$ dữ liệu cung cấp bằng chứng ủng hộ Giả thuyết;
- nếu *p*-value là nhỏ, có nghĩa là dữ liệu không tương thích với Giả thuyết
  $\hookrightarrow$ dữ liệu cung cấp bằng chứng chống lại Giả thuyết
  $\hookrightarrow$ đưa ra quyết định bác bỏ Giả thuyết.

p-value và quyết định

Để xác định mức nhỏ của *p*-value, ta so sánh nó với múc ý nghĩa α:

- nếu *p*-value $\geq \alpha$, thì chưa đủ nhỏ;
- nếu *p*-value $< \alpha$, thì đủ nhỏ để bác bỏ Giả thuyết với sai lầm loại I thấp.

Ví dụ: với *p*-value bằng 0.308:

- so sánh với $\alpha = 0.05 \Longrightarrow$ không bác bỏ Giả thuyết;
- so sánh với $\alpha = 0.1 \Longrightarrow$ không bác bỏ Giả thuyết;
- so sánh với $\alpha = 0.35 \Longrightarrow$ bác bỏ Giả thuyết, với mức sai lầm loại I là 35%;

1 Kiểm định giả thuyết

2 A/B testing hai nhóm

3 A/B testing nhiều nhóm

4 Multi-Armed Bandit

Bài toán

Như đã giới thiệu ở phần trước, A/B test là một thí nghiệm nhằm tìm ra phương pháp A hay phương pháp B là tốt hơn phương pháp còn lại.

$\hookrightarrow$ ta cần so sánh dữ liệu từ hai nhóm A và nhóm B.

- Khi dữ liệu ở dạng biến định lượng $\hookrightarrow$ so sánh giá trị **trung bình** của dữ liệu của hai nhóm $\hookrightarrow$ bài toán kiểm định hai trung bình.
- Khi dữ liệu ở dạng biến định tính $\hookrightarrow$ so sánh giá trị **tỷ lệ** của dữ liệu của hai nhóm $\hookrightarrow$ bài toán kiểm định hai tỷ lệ.

Ví dụ 1: Xét câu hỏi sau: "thời gian tương tác của trang web A và của trang web B có tương đồng nhau?". Nếu:

- thời gian tương tác là không tương đồng
  $\hookrightarrow$ một trong hai trang web (A hoặc B) giúp cải thiện (hoặc giảm) khả năng tương tác với khách hàng;
- thời gian tương tác là như nhau
  $\hookrightarrow$ không có sự cải thiện (giảm) sự tương tác.

So sánh trung bình thời gian tương tác của web A và web B.

Bài toán

Ví dụ 2: Xét câu hỏi sau: "sự lựa chọn cánh gà chiên cay (hot wings) có phụ thuộc giới tính khách hàng (nam/nữ)?". Nếu:

- có ảnh hưởng
  $\hookrightarrow$ một trong hai giới (nam hoặc nữ) mua nhiều hơn giới còn lại
- không ảnh hưởng
  $\hookrightarrow$ lượng mua cánh gà chiên cay là như nhau cho cả hai giới.

So sánh trung bình số lượng cánh gà chiên cay được mua bởi khách hàng nam và khách hàng nữ.

Ví dụ 3: Xét câu hỏi sau: "tỷ lệ chuyển đổi sản phẩm của một mặt hàng trên trang thương mại điện tử có khác biệt giữa hai mức giá?". Nếu:

- có sự khác biệt
  $\hookrightarrow$ một trong hai giá (A hoặc B) tạo ra sự hấp dẫn với khách hàng
- nếu không có sự khác biệt $\hookrightarrow$ cả hai giá đều có sự hấp dẫn khách hàng như nhau.

So sánh tỷ lệ bán được sản phẩm của hai mức giá.

Kiểm định một phía, hai phía

Trong thống kê kiểm định nói chung và A/B testing nói riêng, ta luôn có:

- kiểm định hai phía (two sided)
- kiểm định một phía bên trái (left sided)
- kiểm định một phía bên phải (right sided)

Những kiểm định này được xác định theo tuyên bố của đối thuyết.

- $\mu_A \neq \mu_B \implies$ kiểm định hai phía;
- $\mu_A < \mu_B \implies$ kiểm định một phía bên trái;
- $\mu_A > \mu_B \implies$ kiểm định một phía bên phải.

Kiểm định hai phía được sử dụng khi ta không xác định được hướng so sánh từ ban đầu.

Nếu xác định được hướng/mục tiêu so sánh ngay từ đầu:

- trung bình thời gian tương tác của web A là thấp hơn so với web B;
- trung bình lượng cánh gà được mua bởi nam giới (nhóm A) là cao hơn bởi nữ giới (nhóm B);

thì kiểm định một phía sẽ được áp dụng.

Cách tiếp cận cổ điển

Như đã học ở trong Lý Thuyết Thống Kê hoặc Xác suất - Thống kê, ta sẽ sử dụng

- thống kê

$$T = \frac{(\overline{Y}_A - \overline{Y}_B) - (\mu_A - \mu_B)}{s_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{n_A+n_B-2},$$

  $s_p = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$, với giả định về phân phối chuẩn và phương sai của hai nhóm là đồng nhất, khi kiểm định Giả thuyết "hai trung bình là bằng nhau";

- thống kê

$$Z = \frac{\widehat{p}_A - \widehat{p}_B}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \sim N(0,1),$$

  $\widehat{p} = \frac{n_A\widehat{p}_A + n_B\widehat{p}_B}{n_A + n_B}$, khi kiểm định Giả thuyết "hai tỷ lệ là bằng nhau".

Permutation test

Xét bài toán kiểm định trung bình cho hai nhóm.

Nếu ta sử dụng cách tiếp cận cổ điển, ta cần:

- thông tin về phân phối của dữ liệu trong hai nhóm;
- giả định về sự đồng nhất phương sai giữa hai nhóm;
- cỡ mẫu tương đối lớn.

Tuy nhiên, trong thực tế, đối với khoa học dữ liệu, các điều kiện 1 và 2 thường:

- khó kiểm chứng, hoặc
- không thể đáp ứng.

Do đó, ta cần một phương pháp kiểm định khác, mà không yêu cầu các điều kiện 1 và 2.

Permutation test

Kiểm định hoán vị - Permutation test

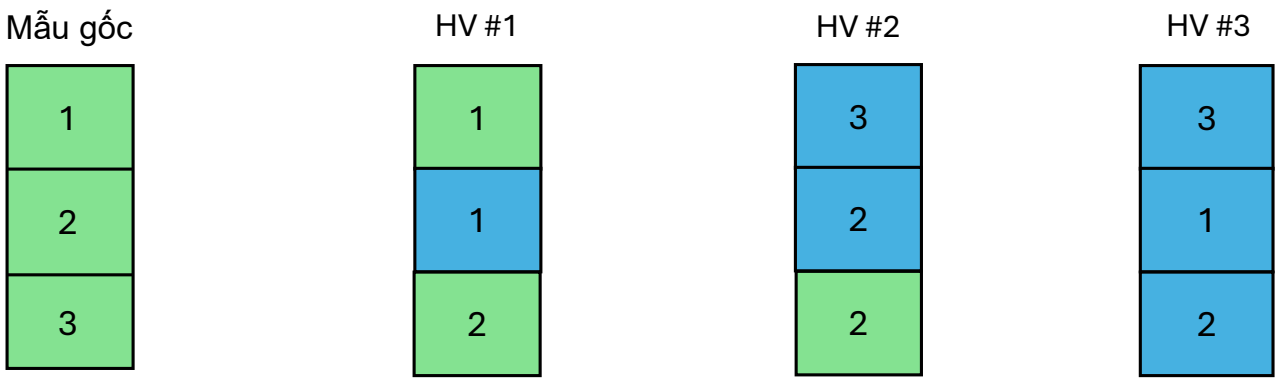
Kiểm định hoán vị (**Permutation test** hay **randomization test**) là một phương pháp kiểm định phi tham số, hoạt động dựa trên hai nguyên tắc:

- Giả định về sự giống nhau giữa hai nhóm;
- xáo trộn và hoán đổi tất cả dữ liệu ở hai nhóm với nhau;

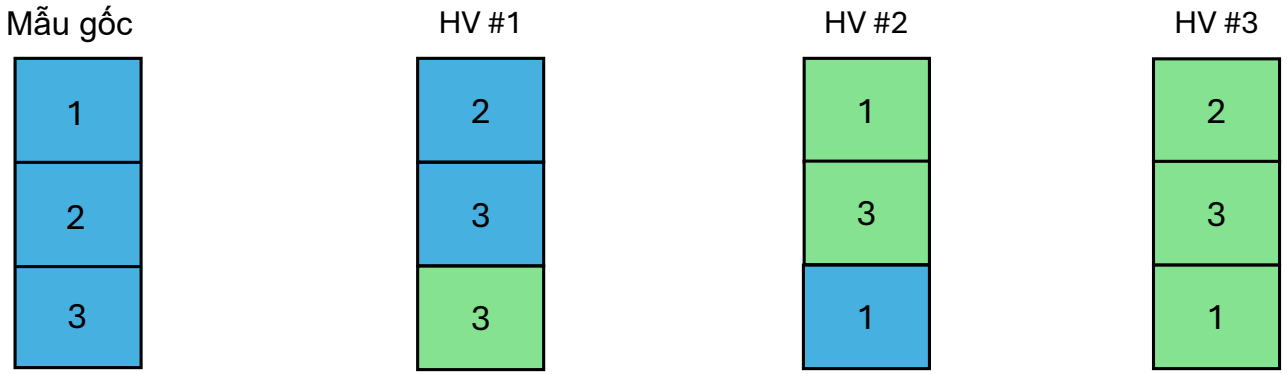
nhằm kiểm tra xem liệu sự khác biệt giữa hai nhóm có phải là do ngẫu nhiên.

Permutation test

Nhóm A



Nhóm B



Permutation test

Kiểm định hoán vị gồm các bước sau:

1. Từ dữ liệu gốc, tính trung bình mẫu $\overline{y}_A$ và $\overline{y}_B$ của hai nhóm, sau đó tính sự khác biệt giữa $\Delta = \overline{y}_A - \overline{y}_B$ (hoặc $\overline{y}_B - \overline{y}_A$, tùy theo đối thuyết).
2. Gộp chung các mẫu lại với nhau, tạo thành một mẫu lớn và xáo trộn chúng.
3. Lấy mẫu ngẫu nhiên không lặp cho nhóm A (với cỡ mẫu như ban đầu) từ mẫu chung.
4. Lấy mẫu ngẫu nhiên không lặp cho nhóm B, từ thành phần còn lại của mẫu chung.
5. Tính trung bình mẫu cho 2 mẫu mới tạo, $\overline{y}_A^*$ và $\overline{y}_B^*$, sau đó, tính sự khác biệt giữa hai trung bình $\overline{y}_A^* - \overline{y}_B^*$ (hoặc $\overline{y}_B^* - \overline{y}_A^*$).
6. Lặp lại bước 3 tới 5 trong R lần, lưu trữ các kết quả của sự khác biệt giữa $\overline{y}_A^* - \overline{y}_B^*$ (hoặc $\overline{y}_B^* - \overline{y}_A^*$).
7. Tính p-value dựa vào mẫu gồm R giá trị $\Delta_i^* = \overline{y}_{A,i}^* - \overline{y}_{B,i}^*$:
   - $p\text{-value} = \frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(|\Delta_i^*| > |\Delta|)$ - kiểm định hai phía;
   - $p\text{-value} = \frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(\Delta_i^* > \Delta)$ - kiểm định một phía bên phải;
   - $p\text{-value} = \frac{1}{R}\sum_{i=1}^{R} \mathrm{I}(\Delta_i^* < \Delta)$ - kiểm định một phía bên trái;

Ví dụ: Web stickiness

Mục tiêu

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Ví dụ: Web stickiness

Mục tiêu

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Các bước xác định mẫu:

Bán hàng tốt hơn $\Longleftrightarrow$ doanh số bán hàng của dịch vụ từ hai trang web có sự khác biệt lớn.
$\hookrightarrow$ ta cần thu thập tổng doanh số bán hàng của hai trang web.

Ví dụ: Web stickiness

Mục tiêu

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Các bước xác định mẫu:

Bán hàng tốt hơn $\Longleftrightarrow$ doanh số bán hàng của dịch vụ từ hai trang web có sự khác biệt lớn.
$\hookrightarrow$ ta cần thu thập tổng doanh số bán hàng của hai trang web.

Tuy nhiên,

- giá trị dịch vụ được bán cao;
- việc bán hàng không thường xuyên;
- chu kỳ bán hàng kéo dài

$\Rightarrow$ cần nhiều thời gian để thu thập đủ doanh số để đánh giá;
$\Rightarrow$ không khả thi, khó thực hiện.

Ví dụ: Web stickiness

Mục tiêu

Một công ty bán một dịch vụ có giá trị cao, muốn kiểm tra xem trang web A hoặc B giúp công ty thực hiện công việc bán hàng tốt hơn.

Các bước xác định mẫu:

Bán hàng tốt hơn $\Longleftrightarrow$ doanh số bán hàng của dịch vụ từ hai trang web có sự khác biệt lớn.
$\hookrightarrow$ ta cần thu thập tổng doanh số bán hàng của hai trang web.

Tuy nhiên,

- giá trị dịch vụ được bán cao;
- việc bán hàng không thường xuyên;
- chu kỳ bán hàng kéo dài

$\Rightarrow$ cần nhiều thời gian để thu thập đủ doanh số để đánh giá;
$\Rightarrow$ không khả thi, khó thực hiện.

Cần thu thập một dữ liệu khác (biến khác), thay thế tổng doanh số bán hàng.

Ví dụ: Web stickiness

Một biến đại diện tiềm năng cho tổng doanh số bán hàng trên web là

số lần nhấp chuột vào trang đích chi tiết

Như tại sao?

Ví dụ: Web stickiness

Một biến đại diện tiềm năng cho tổng doanh số bán hàng trên web là

số lần nhấp chuột vào trang đích chi tiết

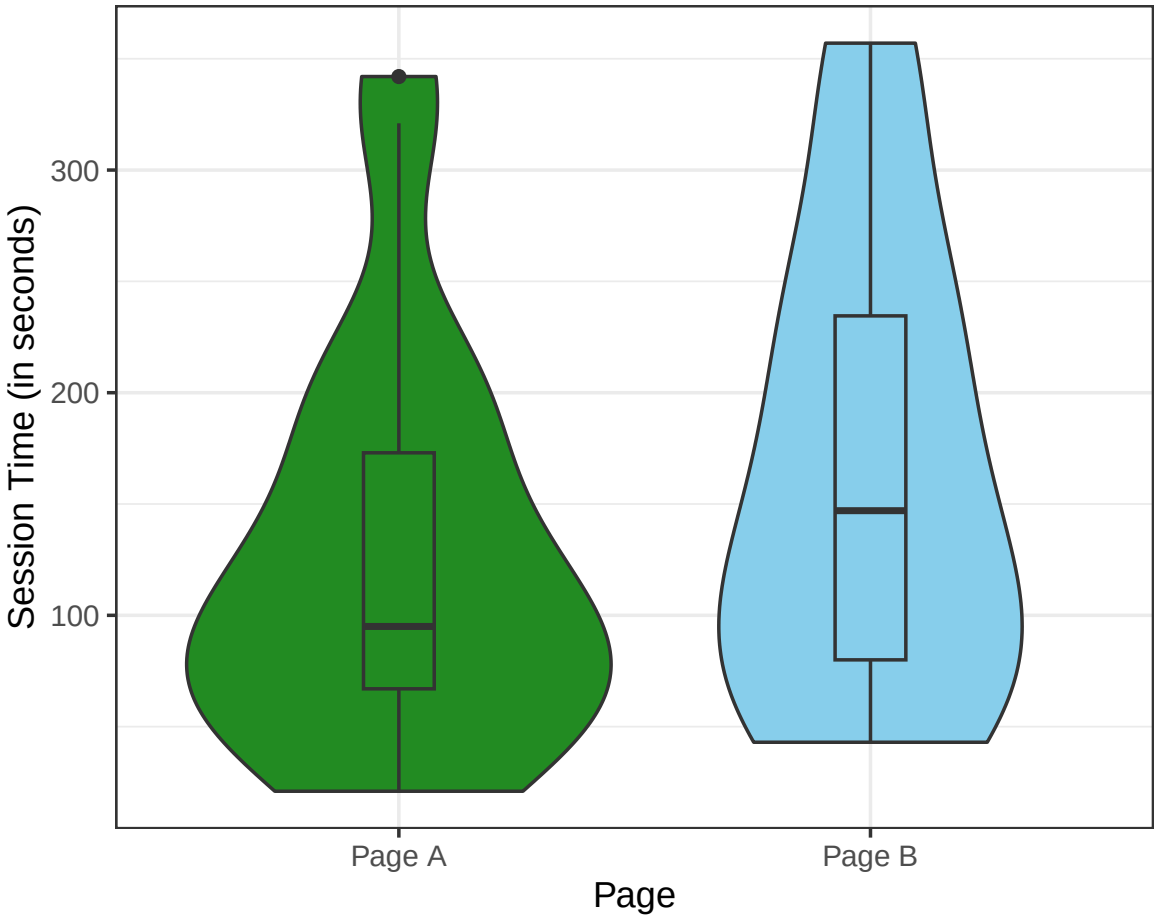
Như tại sao?

Một bản trình bày trên web thu hút sự chú ý của mọi người lâu hơn $\Rightarrow$ doanh số bán hàng cao hơn.

$\hookrightarrow$ dữ liệu là thời gian phiên trung bình (session time), đơn vị giây, so sánh trang web A với trang web B, được ghi nhận bởi Google Analytics.

	Page	Time
1	Page A	0.21
2	Page B	2.53
3	Page A	0.35
4	Page B	0.71
5	Page A	0.67
6	Page B	0.85

Ví dụ: Web stickiness



Có vẻ như thời gian phiên trung bình của trang web B là dài hơn so với trang web A.

Ví dụ: Web stickiness

Ta cần kiểm tra xem liệu thời gian phiên trung bình của trang web B có thực sự dài hơn so với trang web A, hay không.

$\hookrightarrow$ Ta phát biểu Giả thuyết và Đối thuyết như sau:

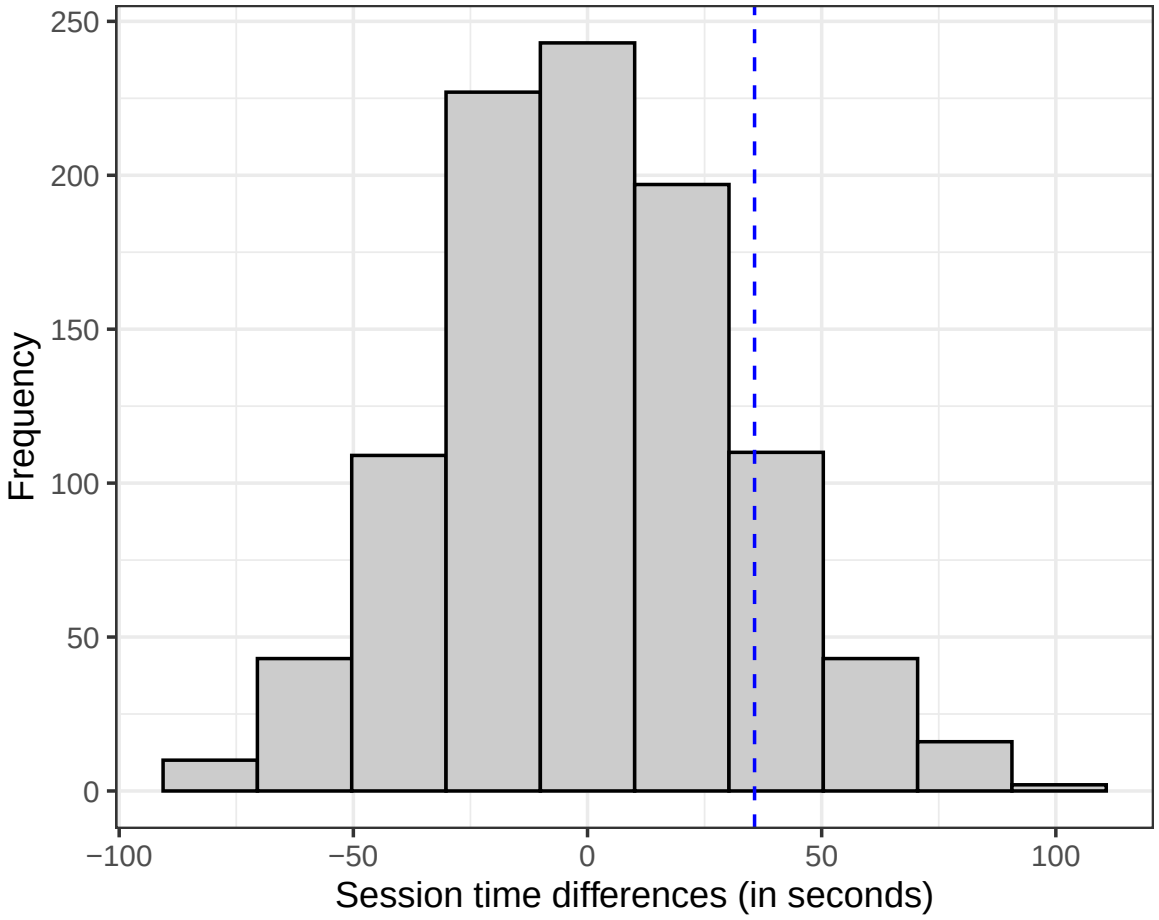
$$\begin{aligned} \text{Giả thuyết:} \quad & \mu_B = \mu_A \\ \text{Đối thuyết:} \quad & \mu_B > \mu_A \end{aligned}$$

Từ dữ liệu ta có một số thông tin

- cỡ mẫu $n_A = 21$, $n_B = 15$;
- trung bình mẫu $\overline{y}_A = 126.33$ giây và $\overline{y}_B = 162$ giây;
- $\Delta = \overline{y}_B - \overline{y}_A = 35.67$.

Ta thực hiện kiểm định hoán vị với $R = 1000$ lần.

Ví dụ: Web stickiness



Ta tính được

$$p\text{-value} = \frac{1}{1000}\sum_{i=1}^{1000} \mathrm{I}(\Delta_i^* > 35.67) = 0.127$$

Ví dụ: Web stickiness

Chọn mức ý nghĩa $\alpha = 0.05$.

$\hookrightarrow$ p-value $> \alpha$

$\hookrightarrow$ ta không đủ cơ sở để bác bỏ Giả thuyết.

$\hookrightarrow$ Sự dài hơn của thời gian phiên trung bình của trang web B so với trang web A là không có **ý nghĩa thống kê** (**statistically significant**).

Kết luận: Có sự cải thiện trong thời phiên trung bình của trang web B so với trang web A. Nhưng sự cải thiện đó không có ý nghĩa thống kê, và có thể là kết quả của sự ngẫu nhiên.

Exhaustive permutation test

Kiểm định hoán vị toàn diện (**Exhaustive permutation test**) là một biến thể của kiểm định hoán vị.

- Thay vì chỉ xáo trộn và chia dữ liệu một cách ngẫu nhiên, kiểm định này xét tất cả các cách chia dữ liệu đó thành 2 nhóm A/B (hoặc nhiều hơn), sau đó xáo trộn và chia ngẫu nhiên dữ liệu vào các nhóm.
- Áp dụng cho các bài toán kiểm định với cỡ mẫu tương đối nhỏ.
- Đôi khi được gọi là kiểm định chính xác (**exact tests**), do đặc tính thống kê của chúng đảm bảo rằng nếu Giả định là đúng thì xác suất sai lầm loại I (α) sẽ luôn được duy trì ở mức ý nghĩa được mong muốn của kiểm định.
- Cách thức xác định *p*-value là tương tự như kiểm định hoán vị.

Ví dụ: với 2 nhóm dữ liệu, mỗi nhóm có lần lượt 9 và 8 quan sát. Lúc này,

- tổng số cách chia dữ liệu (chung) thành 2 nhóm là $2^{17} = 131072$ cách;
- số mẫu có 9 quan sát là $C_{17}^{9} = 24310$;
- số mẫu có 8 quan sát là $C_{17}^{8} = 24310$.

Sau đó, ta sẽ chọn ngẫu nhiên 9 quan sát (trong 17 quan sát) vào nhóm A; phần còn lại thì vào nhóm B.

Bootstrap permutation test

Kiểm định hoán vị bootstrap (**bootstrap permutation test**) là một biến thể của kiểm định hoán vị, trong đó, bước 3 và 4 được thực hiện với nguyên tắc lấy mẫu có hoàn lại.

Bằng cách này, ta có được

- chỉ định ngẫu nhiên cho đối tượng vào nhóm điều trị;
- lựa chọn ngẫu nhiên các đối tượng từ một quần thể.

Power Analysis

Effect size (**cỡ hiệu ứng**) là thước đo về sự chênh lệch giữa trung bình hoặc tỷ lệ giữa hai nhóm, và thông qua kiểm định thống kê, ta muốn phát hiện ra.

Ví dụ:

- trong kiểm định trung bình của hai nhóm, thì effect size là $\Delta = \mu_A - \mu_B$;
- trong kiểm định tỷ lệ của hai nhóm, thì effect size là sai số tương đối giữa hai tỷ lệ $\Delta = \dfrac{p_1 - p_2}{p_2}$.

$\hookrightarrow$ Giả thuyết tương ứng là $\Delta = 0$;

$\hookrightarrow$ Đối thuyết tương ứng là $\Delta \neq 0$.

Độ mạnh của kiểm định (**Power**) là khả năng nhận ra được Giả thuyết là sai, hay tương đương với nhận diện được effect size, tương ứng với một cỡ mẫu cố định.

Power Analysis

- effect size $\Longleftrightarrow$ kích cỡ cục bột cacao;
- cỡ mẫu $\Longleftrightarrow$ độ dày của lưới;
- power $\Longleftrightarrow$ khả năng lọc được cục bột của từng loại lưới.

Power Analysis

Về mặt ý tưởng, cỡ mẫu càng cao, thì độ mạnh của một kiểm định càng lớn.
$\hookrightarrow$ Câu hỏi:

> "Cỡ mẫu bao nhiêu là đủ để đạt được độ mạnh chấp nhận được?"

Trong thống kê, độ mạnh bằng 0.8 (tức 80%) được coi là chấp nhận được.

Để tìm được cỡ mẫu cần thiết, ta cần có:

- effect size mong muốn được nhận ra bởi kiểm định;
- phân phối lý thuyết của dữ liệu (bài toán kiểm định trung bình);
- phương sai hoặc độ lệch chuẩn lý thuyết (bài toán kiểm định trung bình);
- độ mạnh mong muốn của kiểm định;
- kiểm định hai phía, hay kiểm định một phía.

Một quá trình ước tính độ mạnh theo từng kịch bản của cỡ mẫu và effect size được gọi là **phân tích độ mạnh** (**power analysis**).

Phân tích độ mạnh cho kiểm định trung bình thường phức tạp hơn so với kiểm định tỷ lệ.

Power Analysis

Ví dụ: Xét kết quả về tỷ lệ chuyển đổi sản phẩm của một mặt hàng trên trang thương mại điện tử với hai mức giá A và B:

	Giá A	Giá B
Chuyển đổi	200	182
Không chuyển đổi	23539	22406

Ta tính được

- $\widehat{p}_A = 0.8425\%$
- $\widehat{p}_B = 0.8057\%$
- tỷ lệ chênh lệch tương đối là $\dfrac{\widehat{p}_A - \widehat{p}_B}{\widehat{p}_B} \approx 4.57\%$, gần 5%, đủ lớn để có ý nghĩa trong một doanh nghiệp có khối lượng lớn.

Tuy nhiên, ta vẫn có thể hoài nghi về sự chênh lệch này có thực sự có nghĩa hay chỉ là kết quả của sự ngẫu nhiên.

Giả thuyết: "Nếu hai mức giá có cùng tỷ lệ chuyển đổi, liệu sự thay đổi ngẫu nhiên có thể tạo ra sự khác biệt lớn tới 5% không?"

Power Analysis

Nếu Giả thuyết là sai, thì với cỡ mẫu ta đang có (23739 và 22588), độ mạnh của kiểm định là bao nhiêu khi nhận diện effect size 5%?

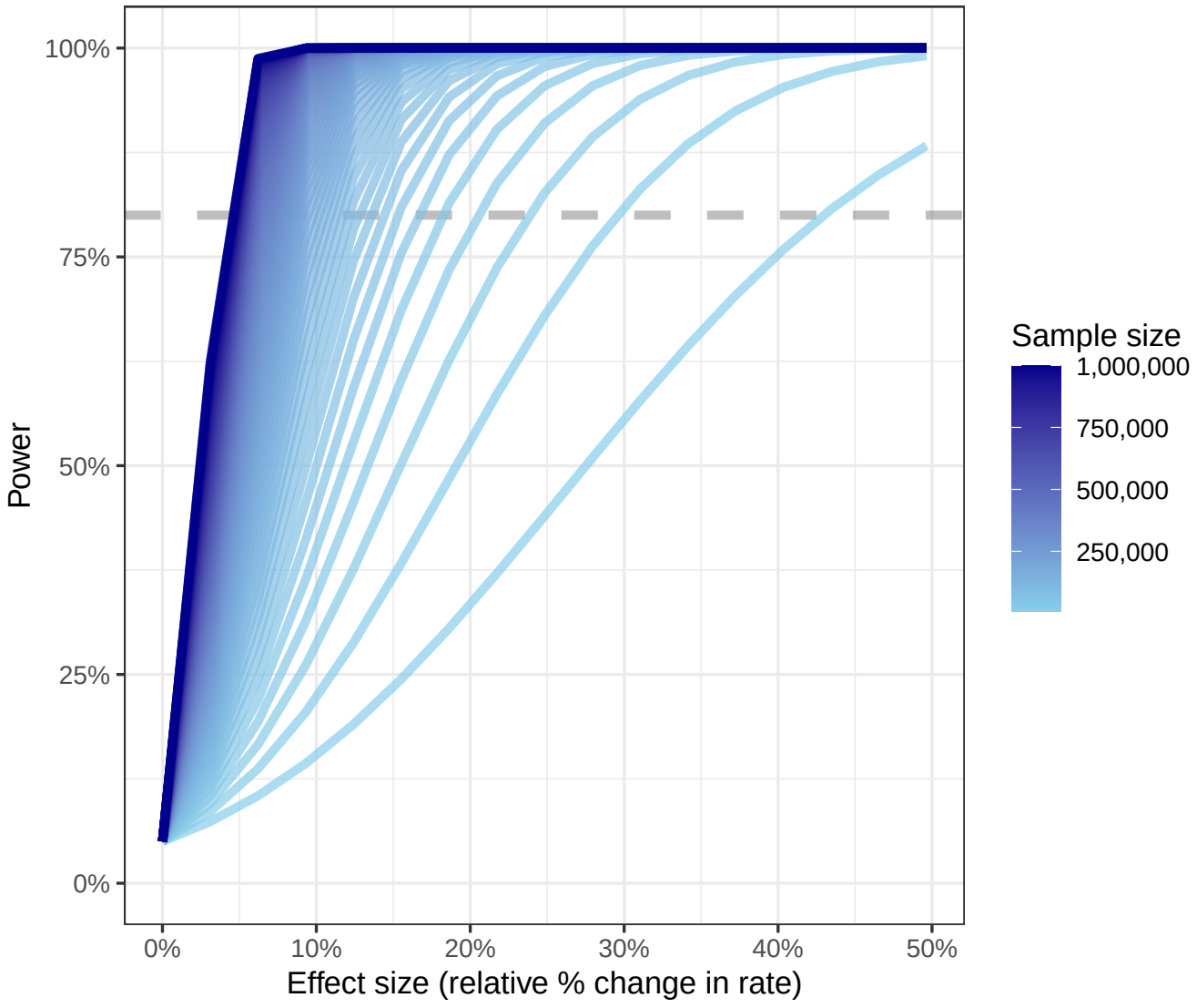
Áp dụng phân tích độ mạnh cho kiểm định tỷ lệ, độ mạnh $\approx$ **11%**.

Hay có nghĩa xác suất mắc sai lầm loại II $\approx$ **89%** (chấp nhận Giả thuyết khi nó thực sự sai).

Và để đạt được độ mạnh xấp xỉ 80%, ta cần cỡ mẫu $\approx$ **746257** cho mỗi mức giá.

Tức gấp gần **32 lần** cỡ mẫu hiện tại.

Power Analysis



1 Kiểm định giả thuyết

2 A/B testing hai nhóm

3 A/B testing nhiều nhóm

4 Multi-Armed Bandit

Bài toán

Giả sử rằng, ta có một bộ dữ liệu về thời gian phiên trung bình của 4 trang web 1, 2, 3 và 4. Ta quan tâm tới câu hỏi:

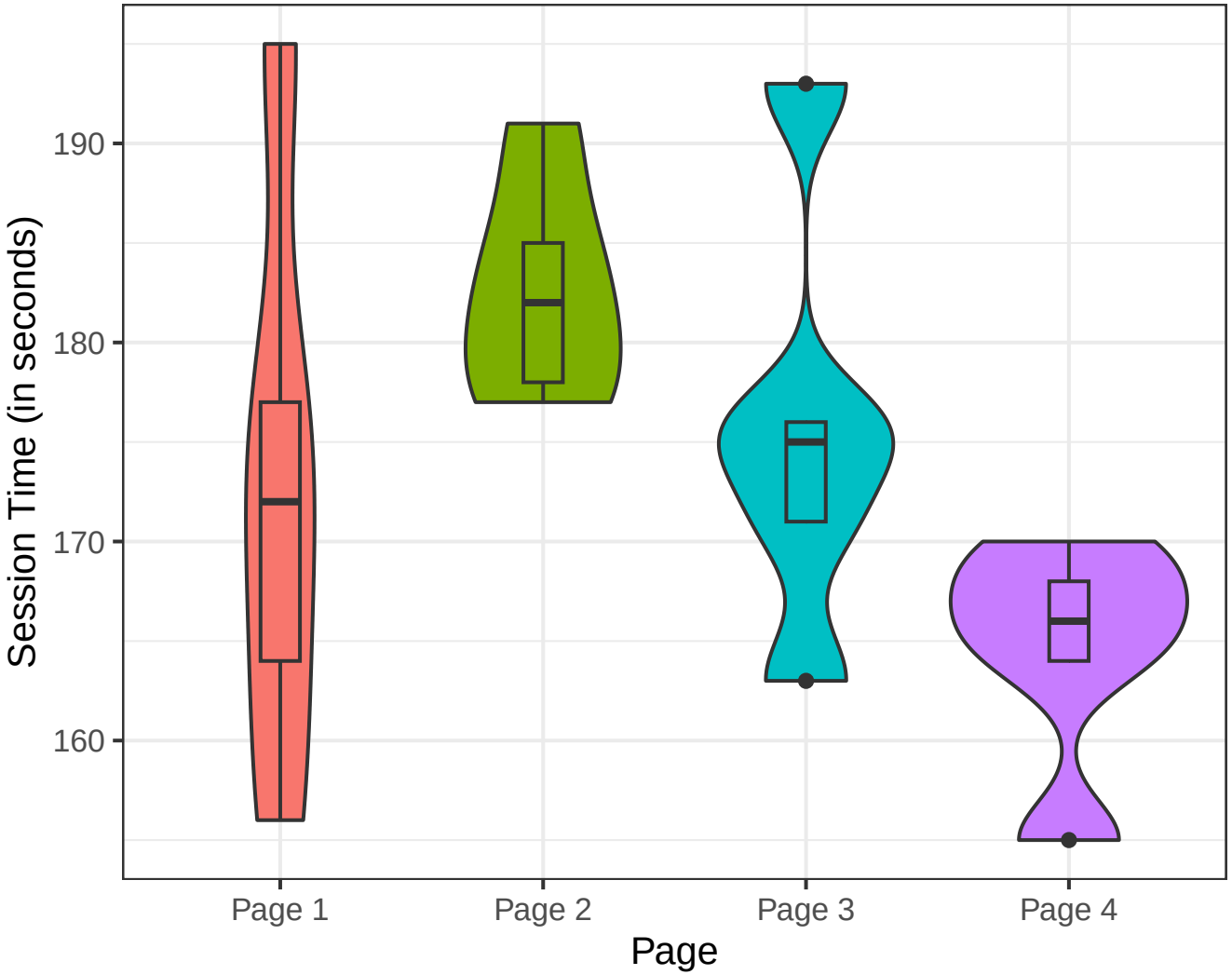
"Thời gian phiên trung bình của 4 trang web là như nhau?"

Lúc này, Giả thuyết cần kiểm chứng sẽ là

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

tức là, ta cần so sánh 4 trung bình cùng một lúc thay vì 2.

Bài toán


Session Time (in seconds)
190
180
170
160
Page 1
Page 2
Page 3
Page 4
Page

Multiple testing

Một cách trực quan, ta sẽ tiến hành so sánh theo từng cặp:

- $\mu_1 = \mu_2$
- $\mu_1 = \mu_3$
- $\mu_1 = \mu_4$
- $\mu_2 = \mu_3$
- $\mu_2 = \mu_4$
- $\mu_3 = \mu_4$

$\hookrightarrow$ tổng cộng 6 cặp.

$\hookrightarrow$ ta cần lặp lại quá trình kiểm định hoán vị trong 6 lần!

Multiple testing

Một cách trực quan, ta sẽ tiến hành so sánh theo từng cặp:

- $\mu_1 = \mu_2$
- $\mu_1 = \mu_3$
- $\mu_1 = \mu_4$
- $\mu_2 = \mu_3$
- $\mu_2 = \mu_4$
- $\mu_3 = \mu_4$

$\hookrightarrow$ tổng cộng 6 cặp.

$\hookrightarrow$ ta cần lặp lại quá trình kiểm định hoán vị trong 6 lần!

Có hai vấn đề nảy sinh:

- thời gian tính toán;
- xác suất tạo ra sai lầm loại I tăng lên.

Multiple testing

Khi số nhóm cần so sánh là k thì ta cần kiểm định $k(k-1)/2$ cặp,

$\hookrightarrow$ Tức là lặp đi lặp lại $k(k-1)/2$ kiểm định hoán vị.

Ví dụ:

- với $k = 10$ thì cần kiểm định 45 Giả thuyết;
- với $k = 100$ thì cần kiểm định 450 Giả thuyết.

Khi ta phải kiểm định m Giả thuyết cùng một lúc, ta có thể có nhiều hơn một lần phạm sai lầm loại I.

Quyết định	Sự thật của Giả thuyết: Đúng	Sai	Tổng
Không bác bỏ	U	W	$m - R$
Bác bỏ	V	S	R
Tổng	m_0	$m - m_0$	m

Khi này, ta quan tâm tới xác suất $\Pr(V \geq 1)$.

Xác suất này được gọi là **family-wise error rate** (FWER).

Multiple testing

Bằng việc giả sử rằng

- m Giả định là đúng;
- m kiểm định là độc lập;
- đặt mức sai lầm loại I cho mỗi kiểm định là α

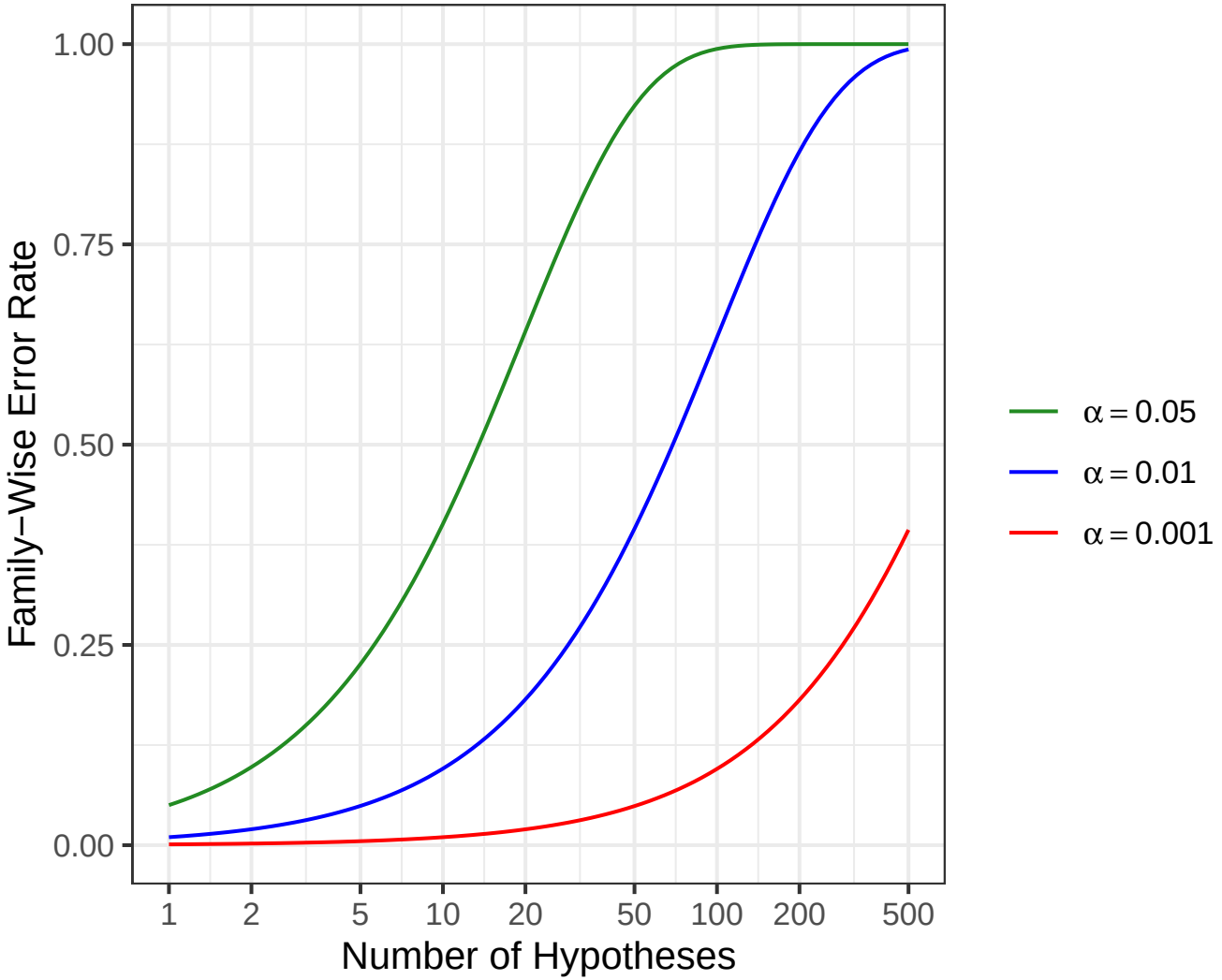
ta xác định được rằng

$$FWER(\alpha) = 1 - (1 - \alpha)^m.$$

Với $\alpha = 0.5$, nếu

- $m = 1$, thì $FWER(\alpha) = 0.05$, chính là mức kiểm soát sai lầm I;
- $m = 6$, thì $FWER(\alpha) = 0.27$;
- $m = 90$, thì $FWER(\alpha) = 0.99$, tức là chắc chắn sẽ có ít nhất 1 sai lầm loại I.

Multiple testing



Multiple testing

Để điều khiển FWER khi kiểm định m Giả thuyết, ta có một số phương pháp hiệu chỉnh:

- Phương pháp Bonferroni (**Bonferroni method**): phương pháp này đề xuất ngưỡng xác suất sai lầm loại I cho từng kiểm định là α/m;
  $\hookrightarrow$ khi đó, ta bác bỏ tất cả Giả thuyết có p-value $< \alpha/m$;
  $\hookrightarrow$ làm như vậy
  $$FWER(\alpha) \leq m \times \frac{\alpha}{m} = \alpha$$
  Ví dụ, với $m = 100$ và $\alpha = 0.05$, khi đó, p-value sẽ so sánh với $0.05/100 = 0.0005$.
- Phương pháp Holm (**Holm's method**)
- Phương pháp Tukey (**Tukey's method**)
- Phương pháp Scheffé (**Scheffé's method**)

ANOVA

Lời giải tốt nhất cho việc kiểm định Giả thuyết

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_k$$

là sử dụng phương pháp **analysis of variance**, hay **ANOVA**.

Phương pháp này sẽ so sánh đồng thời trung bình của nhiều nhóm độc lập trong một phân tích duy nhất.

Phân tích phương sai có vẻ như là một cách gọi sai, do mục đích của chúng ta là so sánh *trung bình*, nhưng việc kiểm tra sự biến động giữa các nhóm cũng tương đương với việc hỏi liệu các trung bình có khác nhau hay không.

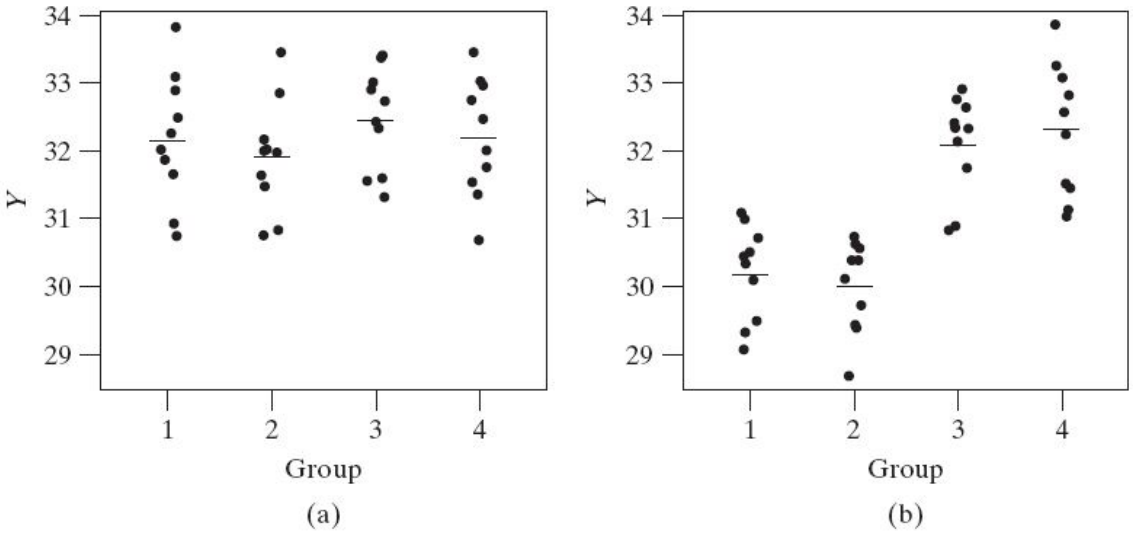
ANOVA được phát triển bởi nhà sinh vật học và nhà thống kê Sir. Ronald Aylmer Fisher[1] (1890 - 1962).

[1] https://en.wikipedia.org/wiki/Ronald_Fisher

ANOVA

Giả sử, ta muốn kiểm định Giả thuyết: $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

Biểu đồ dưới đây minh họa giá trị quan sát của biến Y tương ứng trong 4 nhóm.

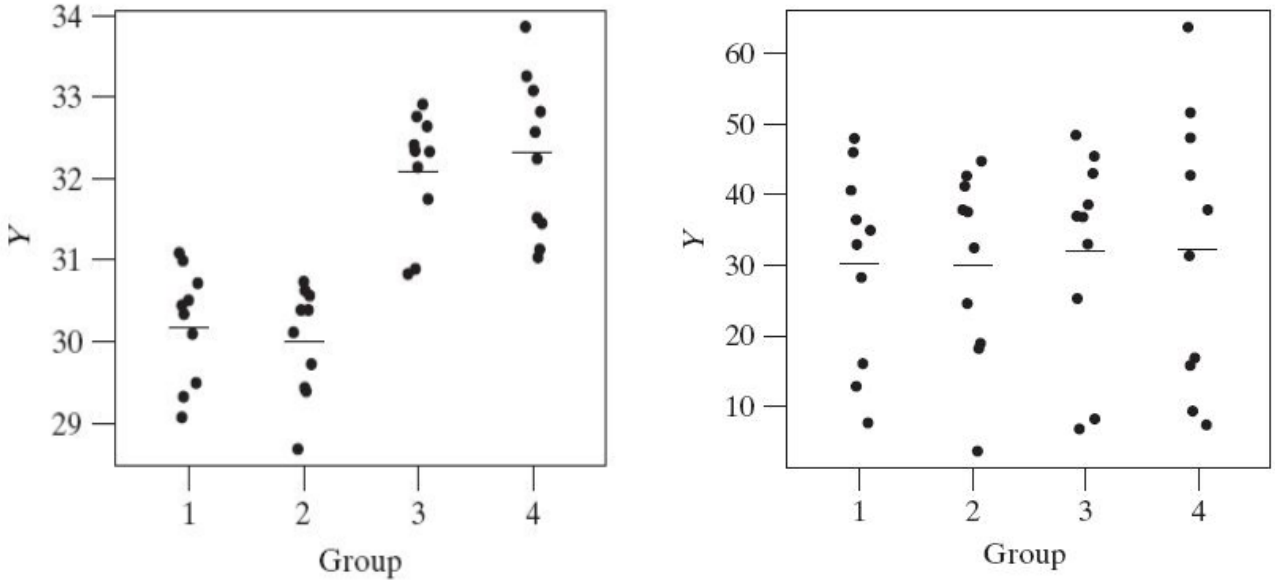


Các trung bình mẫu được hiển thị bởi đường nét liền trong từng nhóm.

- Hình (a), trực giác của chúng ta gợi ý rằng Giả thuyết có lẽ là đúng;
- Hình (b), rõ ràng rằng Giả thuyết là sai, ta có thể nhìn thấy sự khác biệt đáng kể giữa các trung bình của nhóm.

ANOVA

Tuy nhiên,



biểu đồ bên phải thể hiện một trường hợp kém rõ ràng khác.

- Thoạt đầu, ta nghĩ rằng Giả thuyết có lẽ đã đúng.
- Tuy nhiên, thực tế Giả thuyết là sai trong trường hợp - giá trị trung bình mẫu là giống như trong biểu đồ bên trái (tức là biều đồ (b) ở slide trước đó).

Vấn đề ở đây là khi độ lệch chuẩn trong mỗi nhóm là tương đối lớn, nó sẽ khiến cho việc nhận định sự khác biệt của giá trị trung bình trở nên khó khăn.

ANOVA

Để tìm ra bằng chứng thuyết phục về sự khác biệt về trung bình, chúng ta cần tính đến:

(1) sự biến động giữa các nhóm với nhau - variation between groups;

(2) sự biến động vốn có trong các nhóm - variation within groups.

Thông qua việc so sánh độ lớn tương đối của hai loại biến động này - tức là "phân tích phương sai" - mà chúng ta có thể suy luận về trung bình.

Về mặt nguyên tắc, nếu

- sự biến động giữa các nhóm là khá giống với sự biến động trong các nhóm, điều đó có nghĩa là tất cả các nhóm được lấy mẫu từ cùng một quần thể, và do đó, tất cả các trung bình của nhóm đều giống nhau;
- ngược lại, có ít nhất một trung bình là khác với những cái còn lại.

ANOVA

Ta có:

one-way ANOVA so sánh trung bình của ba nhóm trở lên; thuật ngữ "one-way" đề cập đến thực tế là có một biến phân loại xác định các nhóm hoặc các phương pháp điều trị.

two-way ANOVA xem xét ảnh hưởng của hai biến định tính độc lập lên một biến định lượng.

k-way ANOVA xem xét ảnh hưởng của k biến định tính độc lập lên một biến định lượng.

ANOVA

Giả thuyết thống kê

Đối với kiểm định one-way ANOVA, với k nhóm, ta quan tâm tới:

$$\begin{cases} \text{Giả thuyết : } \mu_1 = \mu_2 = \ldots = \mu_k, \\ \text{Đối thuyết : Ít nhất có một trung bình là khác với những cái còn lại.} \end{cases}$$

Điều kiện cho one-way ANOVA

Để thực hiện được kiểm định one-way ANOVA, các điều kiện dưới đây phải được thỏa mãn:

1. Dữ liệu phải được lấy mẫu ngẫu nhiên.
2. Các nhóm phải độc lập với nhau.
3. Dữ liệu trong mỗi nhóm phải tuân theo phân phối chuẩn.
4. Phương sai trong các nhóm phải giống nhau (giả định về sự đồng nhất phương sai - the assumption of homogeneity of variance), tức là

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2.$$

$\hookrightarrow$ ta cần kiểm chứng toàn bộ điều kiện này trước khi phân tích ANOVA.

Permutation ANOVA

Một phương pháp phân tích ANOVA khác, không phụ thuộc vào các giả định trên, đó là sử dụng ý tưởng của permutation test.

1. Gộp chung các mẫu lại với nhau, tạo thành một mẫu lớn và xáo trộn chúng.
2. Lấy mẫu ngẫu nhiên không lặp cho các nhóm j, $j = 1, 2, \ldots, k$ (với cỡ mẫu như ban đầu) từ mẫu chung.
3. Tính trung bình mẫu cho các mẫu mới tạo.
4. Tính phương sai mẫu cho các mẫu mới tạo.
5. Lặp lại các bước từ 2 tới 4, trong R lần.

1 Kiểm định giả thuyết

2 A/B testing hai nhóm

3 A/B testing nhiều nhóm

4 Multi-Armed Bandit

Giới thiệu

Trong các phần trước, ta đã làm quen với A/B testing, và biết rằng nó được sử dụng để kiểm tra:

- sự khác biệt giữa A-B (control - treatment);
- sự khác biệt giữa A-B-C-D

không phải là do ngẫu nhiên.

Tuy nhiên, một nhà khoa học dữ liệu không chỉ dừng lại ở việc kiểm tra này, mà còn cần trả lời câu hỏi:

"Phương pháp nào tốt nhất?"

hoặc câu hỏi

"Nên lựa chọn phương pháp nào?"

Giới thiệu

Ví dụ: Ta có ba máy chơi game, với tỷ lệ thắng cược lần lượt là

- Máy A: 10 chiến thắng trong 50 lượt chơi;
- Máy B: 2 chiến thắng trong 50 lượt chơi;
- Máy C: 4 chiến thắng trong 50 lượt chơi.

Trực quan, ta dễ dàng nói rằng nên tiếp tục chơi máy A.

Tuy nhiên, quyết định như vậy có vội vàng, khi mà ta mới thực hiện 50 lượt chơi?

Điều gì sẽ xảy ra nếu như ta tiếp chơi máy A trong 100 lượt tiếp?

Nếu ta tiếp tục chơi máy A, và số ván thua bắt đầu nhiều lên, liệu ta có tiếc nuối máy C và B?

Có nên thử với máy C? $\hookrightarrow$ điều này dẫn tới nhu cầu về một thuật toán mô phỏng tiến trình chơi của 3 máy, để nhìn thấy cơ hội ở máy nào là rõ rệt hơn.

$\hookrightarrow$ Multi-Armed Bandit Algorithm.

Khai thác và khám phá

Multi-Armed Bandit Algorithm liên quan tới hai khái niệm

- khai thác (exploitation): chọn những lựa chọn có vẻ tốt nhất dựa trên kết quả trong quá khứ
- khám phá (exploration): chọn các lựa chọn chưa được thử (hoặc chưa thử đủ)

Khai thác có liên quan những khái niệm:

- "tham lam" (greedy) và "thiển cận" (short-sighted).
- khai thác quá nhiều $\Longleftrightarrow$ tiếc nuối vì đã bỏ lỡ những cơ hội ngon ăn khác chưa được khám phá.

Khám phá liên quan tới

- "thu thập thông tin" và "nhìn xa trông rộng" (long-sighted).
- khám phá quá nhiều $\Longleftrightarrow$ hối tiếc vì đã lãng phí thời gian vào những việc "ngớ ngẩn".

$\hookrightarrow$ Làm thế nào để cân bằng giữa Khai thác và Khám phá để kết hợp giữa lợi ích thông tin và lợi nhuận một cách tối ưu nhất.

Khai thác và khám phá

Ví dụ:

- Lựa chọn nhà hàng
  - **Khai thác**: Đi tới nhà hàng yêu thích
  - **Khám phá**: Tới một cửa hàng mới
- Biểu ngữ quảng cáo online
  - **Khai thác**: Hiển thị dòng quảng cáo thành công nhất
  - **Khám phá**: Hiển thị một dòng quảng cáo mới
- Lựa chọn môn học
  - **Khai thác**: Lựa chọn những môn được cho là dễ học và điểm cao
  - **Khám phá**: Lựa chọn môn mới/giảng viên mới
- Lựa chọn phim
  - **Khai thác**: Xem những phim/dòng phim mà mình thích
  - **Khám phá**: Thử những phim/dòng phim mới

Các thuật toán Multi-Armed Bandit

Có nhiều thuật toán khác nhau đã được phát triển cho Multi-Armed Bandit:

- epsilon-greedy algorithm
- the softmax algorithm
- the upper confidence bound algorithm
- Multi-Armed Bandit with Thompson Sampling
- Epsilon-First

Mỗi thuật toán có ưu nhược điểm khác nhau.

Ví dụ: lựa chọn quảng cáo

Một khách muốn có lời khuyên nên sử dụng quảng cáo nào trong 5 biến thể.

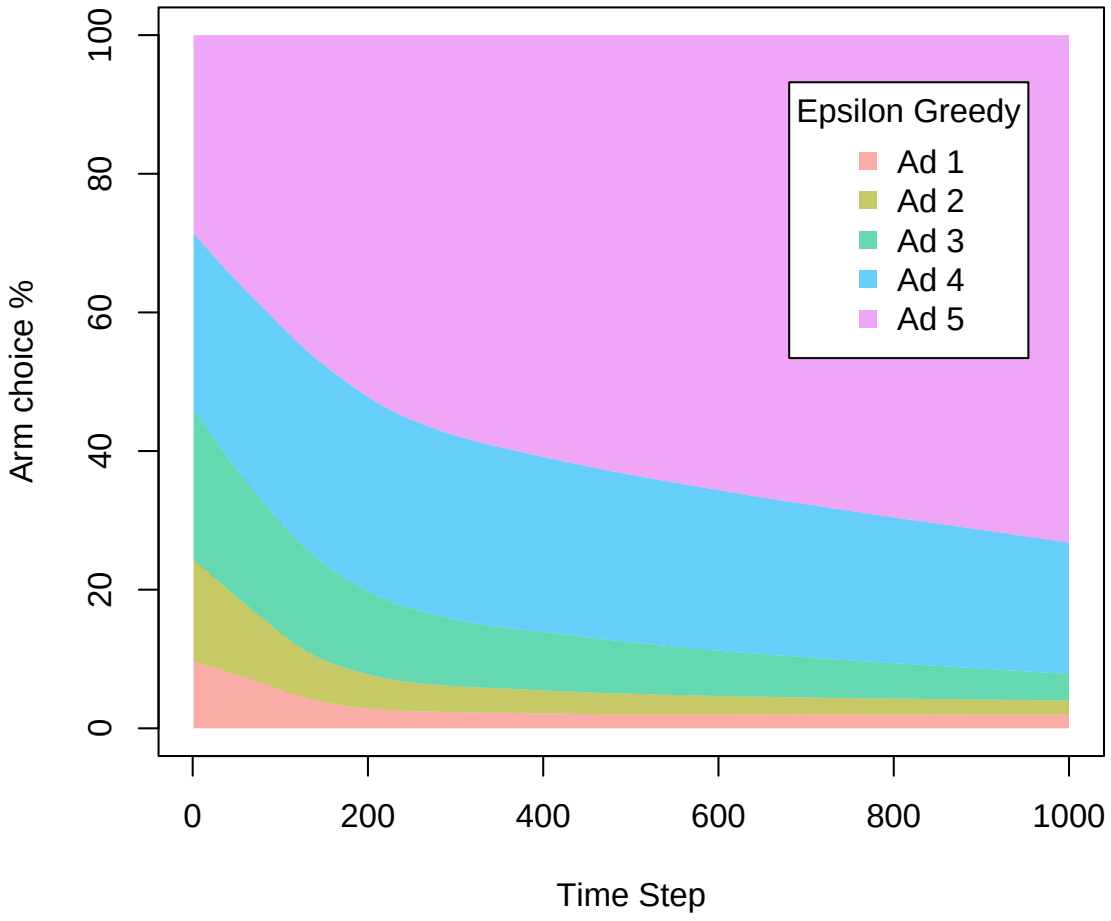
Biết rằng, tỷ lệ chuyển đổi sản phẩm của 5 quảng cáo:

- Quảng cáo 1: 5%
- Quảng cáo 2: 10%
- Quảng cáo 3: 15%
- Quảng cáo 4: 20%
- Quảng cáo 5: 25%

Quảng cáo đóng vai trò là các arms.

Để đưa ra lựa chọn quảng cáo nào là tốt, cũng như không phải hối tiếc những quảng cáo khác, ta áp dụng epsilon-greedy algorithm để mô phỏng % chọn lựa của 5 loại quảng cáo trong 1000 lần thử.

Ví dụ: lựa chọn quảng cáo



Thông qua hình minh họa kết quả, ta thấy rằng Quảng cáo số 5 được lựa chọn nhiều nhất.

Đề tài cuối kỳ

Ta có một số đề tài cuối cùng liên quan tới chủ đề "Multi-Armed Bandit Algorithms".

1. Tìm hiểu thuật toán Epsilon-greedy algorithm và các ứng dụng trong thực tế.
2. Tìm hiểu thuật toán Softmax algorithm và các ứng dụng trong thực tế.
3. Tìm hiểu thuật toán Upper confidence bound algorithm và các ứng dụng trong thực tế.
4. Tìm hiểu thuật toán Multi-Armed Bandit with Thompson Sampling và các ứng dụng trong thực tế.