Bài giảng 3: 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 Tổng quan

2 Hai nhóm độc lập

3 Hai nhóm ghép cặp

4 Bootstrap test một mẫu

5 Bootstrap test cho 2 mẫu

1 Tổng quan

2 Hai nhóm độc lập

3 Hai nhóm ghép cặp

4 Bootstrap test một mẫu

5 Bootstrap test cho 2 mẫu

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.

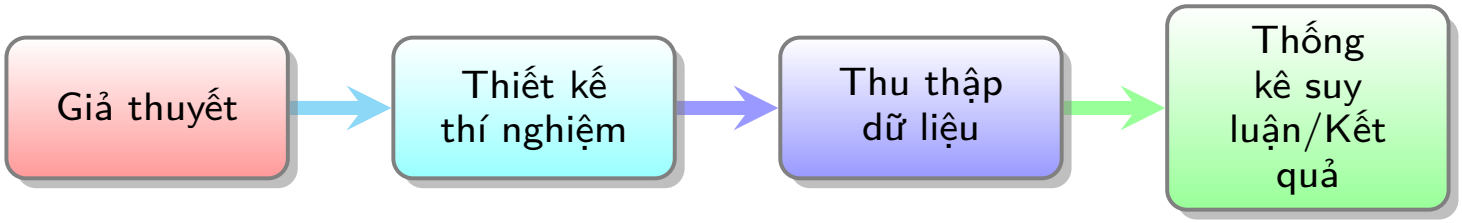
Khái niệm cơ bản

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 không - 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: "Điểm thi trung bình của học sinh được luyện thi và không được luyện thi 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:

(+) "Điểm thi trung bình của học sinh được luyện thi và không được luyện thi là **khác biệt**."

(+) "Điểm thi trung bình của học sinh được luyện thi là **cao hơn** của học sinh không được luyện thi."

(+) "Điểm thi trung bình của học sinh được luyện thi là **thấp hơn** của học sinh không được luyện thi."

Thông qua dữ liệu thu thập từ việc thiết kế thí nghiệm, ta cần tìm bằng chứng thống kê để chứng minh một trong hai giả thuyết là phù hợp.

A/B testing

Một kiểm định A/B (A/B testing) là một thí nghiệm với hai nhóm nhằm xác định xem nhóm nào 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 không và đối thuyết của A/B testing thường là

- Giả thuyết không : 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:

- hiệu quả của các phương pháp điều trị khác nhau;
- hoặc 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).

Ứng dụng của A/B testing

A/B testing được áp dụng phổ biến trong nhiều lĩnh vực:

- Khoa học sự sống (sinh học, môi trường, y tế)
- Khoa học xã hội (tâm lý hành vi)
- Giáo dục (kiểm tra phương pháp giảng dạy, nền tảng giảng dạy - học tập)
- Kinh tế - Tài chính
- Khoa học dữ liệu

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 không H_0 Đúng	Sai
Không bác bỏ H_0	Chính xác	Sai lầm loại II
Bác bỏ H_0	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 không H_0 khi nó 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 không H_0 khi nó thực sự sai.

Sai lầm trong kiểm định

- Xác suất mắc sai lầm loại I thường được ký hiệu là α và nó tương ứng 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 không H_0 là sai.

Trong khuôn khổ của một kiểm định thống kê, ta giả định rằng Giả thuyết không H_0 là đúng. Vì vậy, chúng ta có thể nhận được:

- quyết định chính xác nếu ta không bác bỏ Giả thuyết không H_0; hoặc,
- sai lầm loại I nếu ta bác bỏ Giả thuyết không H_0.

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

Định nghĩa p-value

p-value là xác suất, được tính dưới giả định rằng giả thuyết không H_0 đúng, để thống kê kiểm định nhận giá trị **ít nhất cực đoan** như giá trị quan sát được từ dữ liệu, tức là

$$p\text{-value} = \Pr(T \geq t_{\text{obs}} | H_0 \text{ là đúng}),$$

trong đó, T là thống kê kiểm định và t_{obs} là giá trị quan sát được từ dữ liệu.

p-value và quyết định

Ví dụ:

- p-value $= 0.308$ có nghĩa là:

  dưới giả định rằng Giả thuyết không H_0 đúng, xác suất để thu được giá trị của thống kê kiểm định ít nhất cực đoan như giá trị quan sát được từ dữ liệu là khoảng 30.8%.

  Nói cách khác, nếu H_0 đúng, thì những kết quả như dữ liệu hiện tại là **không phải hiếm gặp**.

- p-value $= 0.031$ có nghĩa là:

  dưới giả định rằng Giả thuyết không H_0 đúng, xác suất để thu được giá trị của thống kê kiểm định cực đoan bằng hoặc cực đoan hơn giá trị quan sát được từ dữ liệu là khoảng 3.1%.

  Nói cách khác, nếu H_0 đúng, thì những kết quả như dữ liệu hiện tại là **hiếm gặp**.

p-value và quyết định

p-value có thể được xem như một thước đo mức độ tương thích giữa dữ liệu quan sát được và giả thuyết không H_0.

- nếu p-value lớn, dữ liệu được xem là tương thích với H_0
  $\hookrightarrow$ dữ liệu chưa cung cấp đủ bằng chứng để bác bỏ H_0;
- nếu p-value nhỏ, dữ liệu được xem là ít tương thích với H_0
  $\hookrightarrow$ dữ liệu cung cấp bằng chứng chống lại H_0;
  $\hookrightarrow$ có thể đưa ra quyết định bác bỏ H_0.

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 không H_0;
- so sánh với $\alpha = 0.1 \Longrightarrow$ không bác bỏ Giả thuyết không H_0;
- so sánh với $\alpha = 0.35 \Longrightarrow$ bác bỏ Giả thuyết không H_0, với mức sai lầm loại I là 35%;

1 Tổng quan

2 Hai nhóm độc lập

3 Hai nhóm ghép cặp

4 Bootstrap test một mẫu

5 Bootstrap test cho 2 mẫu

Bài toán

A/B testing 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 $\Rightarrow$ so sánh giá trị **trung bình** của dữ liệu của hai nhóm $\Rightarrow$ bài toán kiểm định hai trung bình.
- Khi dữ liệu ở dạng biến định tính $\Rightarrow$ so sánh giá trị **tỷ lệ** của dữ liệu của hai nhóm $\Rightarrow$ bài toán kiểm định hai tỷ lệ.

Ví dụ 1: Xét câu hỏi sau: "Điểm thi trung bình của học sinh được luyện thi và không được luyện thi có tương đồng nhau?". Nếu:

- điểm thi là không tương đồng
  $\hookrightarrow$ việc luyện thi **có thể** giúp cải thiện (hoặc giảm) khả năng làm bài thi của học sinh;
- điểm thi là như nhau
  $\hookrightarrow$ không có sự cải thiện (giảm) khả năng làm bài thi của học sinh.

So sánh trung bình điểm thi của nhóm A (không được luyện thi) và nhóm B (được luyện thi).

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 điểm thi của nhóm A là thấp hơn so với nhóm 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

Xét bài toán kiểm định

$$\begin{cases} H_0: & \mu_A - \mu_B = 0, \\ H_1: & \mu_A - \mu_B \neq 0. \end{cases}$$

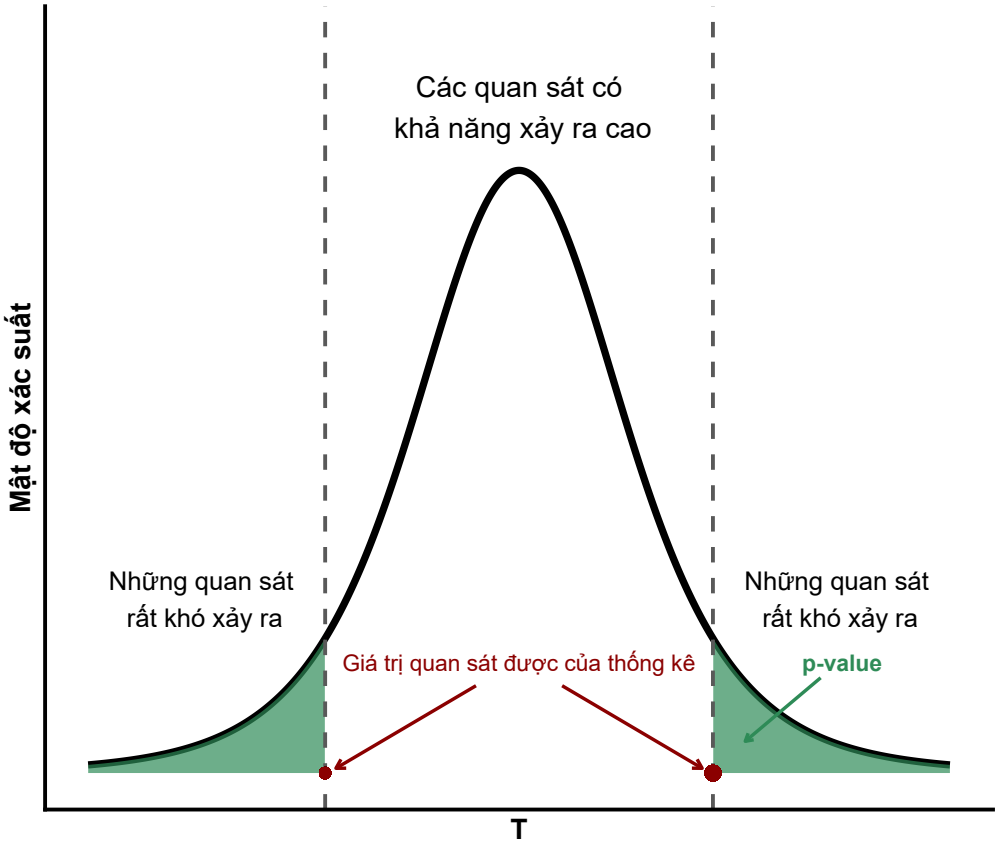
Như đã học ở trong Thống kê cơ bản, ta sẽ sử dụng thống kê

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

$s_p = \dfrac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$, với các giả định:

- dữ liệu của hai nhóm tuân theo phân phối chuẩn;
- phương sai của hai nhóm là đồng nhất.

Cách tiếp cận cổ điển



Ta xác định p-value bởi:

$$p\text{-value} = \Pr\left(|T| \geq |t_{\text{obs}}| \,\middle|\, H_0 \text{ đúng}\right)$$

Cách tiếp cận cổ điển

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.

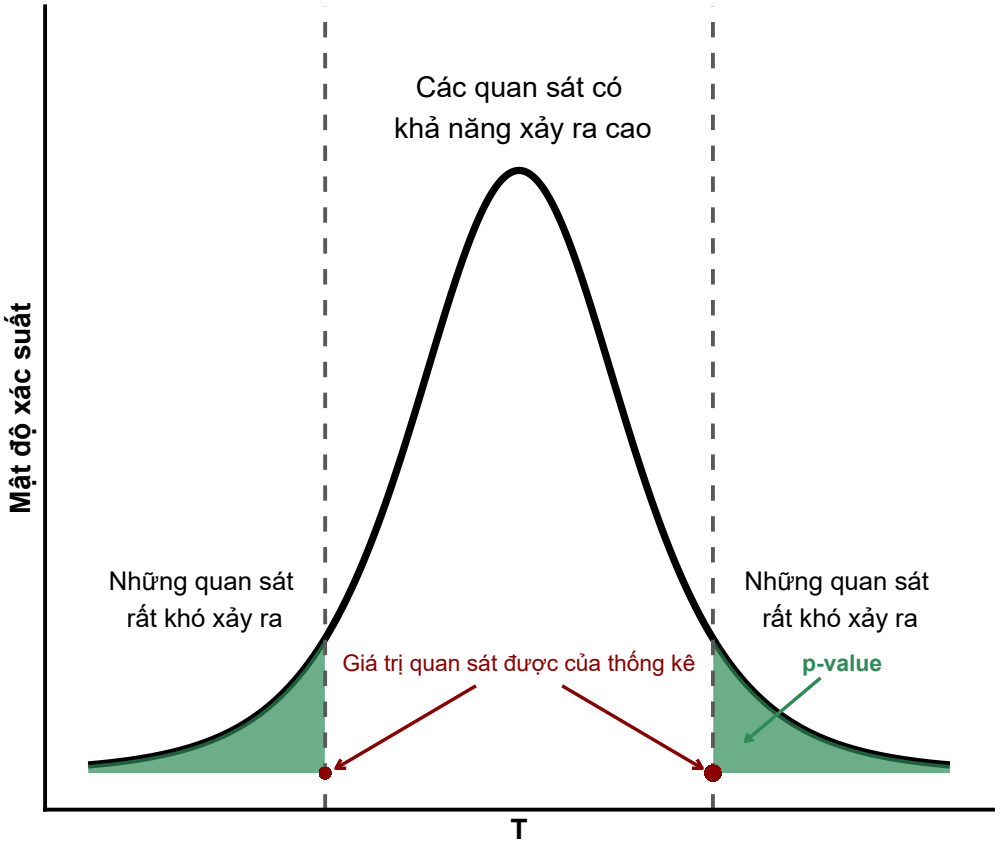
Tuy nhiên, trong thực tế, các điều kiện 1 và 2 thường:

- 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.

Phương pháp resampling

p-value được xác định dựa trên phân phối lý thuyết $F_T(t)$ của T.



Phương pháp resampling

Nếu không yêu cầu các điều kiện 1 và 2:

- thông tin về phân phối của dữ liệu trong hai nhóm;
- giả định về sự đồng nhất phương sai giữa hai nhóm;

$\Rightarrow$ phân phối lý thuyết (t-student) của T là không được xác định.

Tuy nhiên, nếu ta có mẫu ngẫu nhiên $T_1, \ldots, T_m$, với T_j là thống kê của kiểm định ứng với mẫu ngẫu nhiên thứ j (có cỡ mẫu n).

$\Rightarrow$ xác định được phân phối thực nghiệm cho T:

$$\widehat{F}_T(t) = \frac{1}{n} \sum_{j=1}^{m} \mathrm{I}\left(T_j \leq t\right).$$

Theo kết quả lý thuyết, $\widehat{F}_T(t)$ sẽ gần với $F_T(t)$ khi m lớn.

$\Rightarrow$ sử dụng $\widehat{F}_T(t)$ thay cho $F_T(t)$ chưa biết.

Phương pháp resampling

Trong thực tế, ta không thể thực hiện lấy dữ liệu m lần, mà chỉ có 1 lần duy nhất. $\Rightarrow$ cách làm trên không khả thi.

Tuy nhiên, nếu dữ liệu thu được là đủ tốt để đại diện cho quần thể

$\Rightarrow$ ta có thể dùng mẫu đó như quần thể

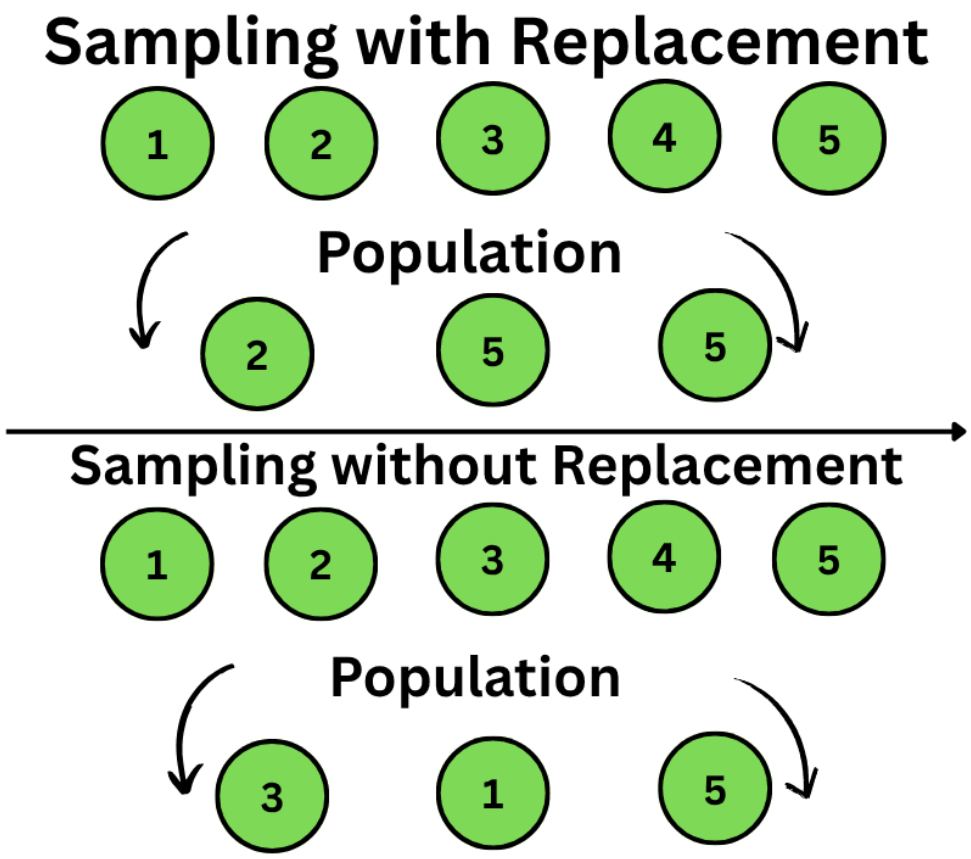
$\Rightarrow$ tạo mẫu ngẫu nhiên từ dữ liệu này.

Khi ta tạo đi tạo lại mẫu ngẫu nhiên từ 1 dữ liệu gốc $\Rightarrow$ ta đang thực hiện phương pháp **resampling**.

Ta có hai cách lấy mẫu chính:

- lấy mẫu không hoàn lại (sampling without replacement);
- lấy mẫu có hoàn lại (sampling with replacement).

Phương pháp resampling



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**.

Trong kiểm định hoán vị, giả thuyết gốc H_0 được phát biểu là phân phối xác suất của hai nhóm là như nhau, tức là $F_{Y_A} = F_{Y_B}$.

$\Rightarrow$ Nếu H_0 đúng thì có nghĩa là dữ liệu có thể hoán đổi (exchangeable) giữa các nhóm.

Definition (Exchangeability)

Xét mẫu ngẫu nhiên $\mathbf{Y} = (Y_1, \ldots, Y_n)$ có hàm mật độ xác suất đồng thời $f_{\mathbf{Y}}(y_1, \ldots, y_n)$. Nếu $f_{\mathbf{Y}}(y_1, \ldots, y_n) = f_{\mathbf{Y}}(y_{\pi_1}, \ldots, y_{\pi_n})$ với mọi hoán vị π của tập chỉ số $\{1, \ldots, n\}$, thì ta nói $Y_1, \ldots, Y_n$ là có thể hoán đổi.

Permutation test

Nhóm A

Mẫu gốc	HV #1	HV #2	HV #3
1	1	3	3
2	1	2	1
3	2	2	2

Nhóm B

Mẫu gốc	HV #1	HV #2	HV #3
1	2	1	2
2	3	3	3
3	3	1	1

Permutation test

Xét hai nhóm dữ liệu

- $(Y_{A,1}, Y_{A,2}, \ldots, Y_{A,n_A})$ độc lập cùng phân phối F_{Y_A};
- $(Y_{B,1}, Y_{B,2}, \ldots, Y_{B,n_B})$ độc lập cùng phân phối F_{Y_B}.

Gộp lại ta có một dữ liệu đồng thời

$$(Y_{A,1}, Y_{A,2}, \ldots, Y_{A,n_A}, Y_{B,1}, Y_{B,2}, \ldots, Y_{B,n_B}).$$

Giả định rằng H_0 là đúng, tức là $F_{Y_A} = F_{Y_B} = F_Y$, khi đó, việc hoán đổi nhãn của hai nhóm dữ liệu cũng sẽ không thay đổi phân phối của chúng, tức là,

$$f_{\mathbf{Y}}(y_1, \ldots, y_n) = f_{\mathbf{Y}}(y_{\pi_1}, \ldots, y_{\pi_n})$$

với mọi hoán vị π của tập chỉ số $\{1, \ldots, n\}$ và $n = n_A + n_B$.
Nhắc lại: phân phối của biến ngẫu nhiên được đặc trưng bởi:

- trung bình (kỳ vọng),
- trung vị (median),
- phương sai (độ phân tán của dữ liệu).

$\Rightarrow$ có thể sử dụng các đặc trưng này như là thống kê kiểm định cho việc so sánh các phân phối của mẫu hoán vị.

Permutation test

Sử dụng sự khác biệt trung bình làm thống kê kiểm định.

$\Rightarrow$ cặp giả thuyết được viết lại thành

$$\begin{cases} H_0: & \mu_A - \mu_B = 0 \\ H_1: & \mu_A - \mu_B \neq 0. \end{cases}$$

Khi này, giá trị thống kê kiểm định trên mẫu sẽ là $t_{\text{diff}} = \overline{Y}_A - \overline{Y}_B$.

Nếu giả thuyết H_0 là đúng thì sẽ rất bất ngờ nếu ta quan sát thấy t_{diff} là lớn hơn hoặc nhỏ hơn hầu hết các giá trị $t_{\text{diff},j}$ (với $j = 1, \ldots, n!$, $n!$ là tổng số lượng hoán vị).

Khi này, ta định nghĩa được p-value là

$$p\text{-value} = \frac{1}{n!} \sum_{j=1}^{n!} \mathrm{I}\left(|t_{\text{diff},j}| \geq |t_{\text{diff}}|\right).$$

Nếu p-value $\leq \alpha \Rightarrow$ giả thuyết gốc H_0 là không tương thích với dữ liệu, hay việc gán nhãn của hai nhóm là **không** ngẫu nhiên

$\Rightarrow$ có sự khác biệt về giá trị của hai nhóm.

Permutation test

Quy trình kiểm định hoán vị được thực hiện qua các bước như sau:

Bước 1 Tính $t_{\text{diff}} = \overline{Y}_A - \overline{Y}_B$ từ dữ liệu gốc.

Bước 2 Gộp chung hai dữ liệu thành một dữ liệu chung, ký hiệu chung là $(Y_1, \ldots, Y_n)$ với $n = n_A + n_B$.

Bước 3 Hoán vị các vị trí trong dữ liệu chung, sau đó, chia ngẫu nhiên dữ liệu thành hai bộ dữ liệu có kích cỡ n_A và n_B.

Bước 4 Tính $t_{\text{diff},\pi} = \overline{Y}_{A,\pi} - \overline{Y}_{B,\pi}$ từ dữ liệu hoán vị.

Bước 5 Lặp lại bước 3 và 4 trong $n!$ lần (tương ứng với tổng số lần hoán vị).

Bước 6 Tính p-value theo công thức và so sánh với mức ý nghĩa α.

Tuy nhiên, $n!$ sẽ rất lớn ngay cả khi với cỡ mẫu nhỏ ($10! = 3,628,800$)

Permutation test

$\Rightarrow$ không thể tính toán được toàn bộ số hoán vị. Vì vậy, trong áp dụng thực tế, bước 5 và 6 sẽ được thay thế bởi:

*Bước 5** Lặp lại bước 3 và 4 trong R lần (tối thiểu 10,000 lần).

*Bước 6** Tính p-value theo công thức

$$p\text{-value} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}\left(|t_{\mathrm{diff},j}| \geq |t_{\mathrm{diff}}|\right),$$

và so sánh với mức ý nghĩa α.

Permutation test

Trong trường hợp ta muốn kiểm định các cặp giả thuyết một phía

$$\begin{cases} H_0 : \mu_A \leq \mu_B \\ H_1 : \mu_A > \mu_B \end{cases}, \qquad \begin{cases} H_0 : \mu_A \geq \mu_B \\ H_1 : \mu_A < \mu_B \end{cases}$$

khi đó, p-value được xác định bởi:

- cặp giả thuyết kiểm định một phía bên phải

$$p\text{-value} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}\left(t_{\mathrm{diff},j} \geq t_{\mathrm{diff}}\right);$$

- cặp giả thuyết kiểm định một phía bên trái

$$p\text{-value} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}\left(t_{\mathrm{diff},j} \leq t_{\mathrm{diff}}\right).$$

Ví dụ mô phỏng: `https://istats.shinyapps.io/PermDist_2samples/`

Ví dụ: Tăng tính tương tác tới trang web

Bài toán

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

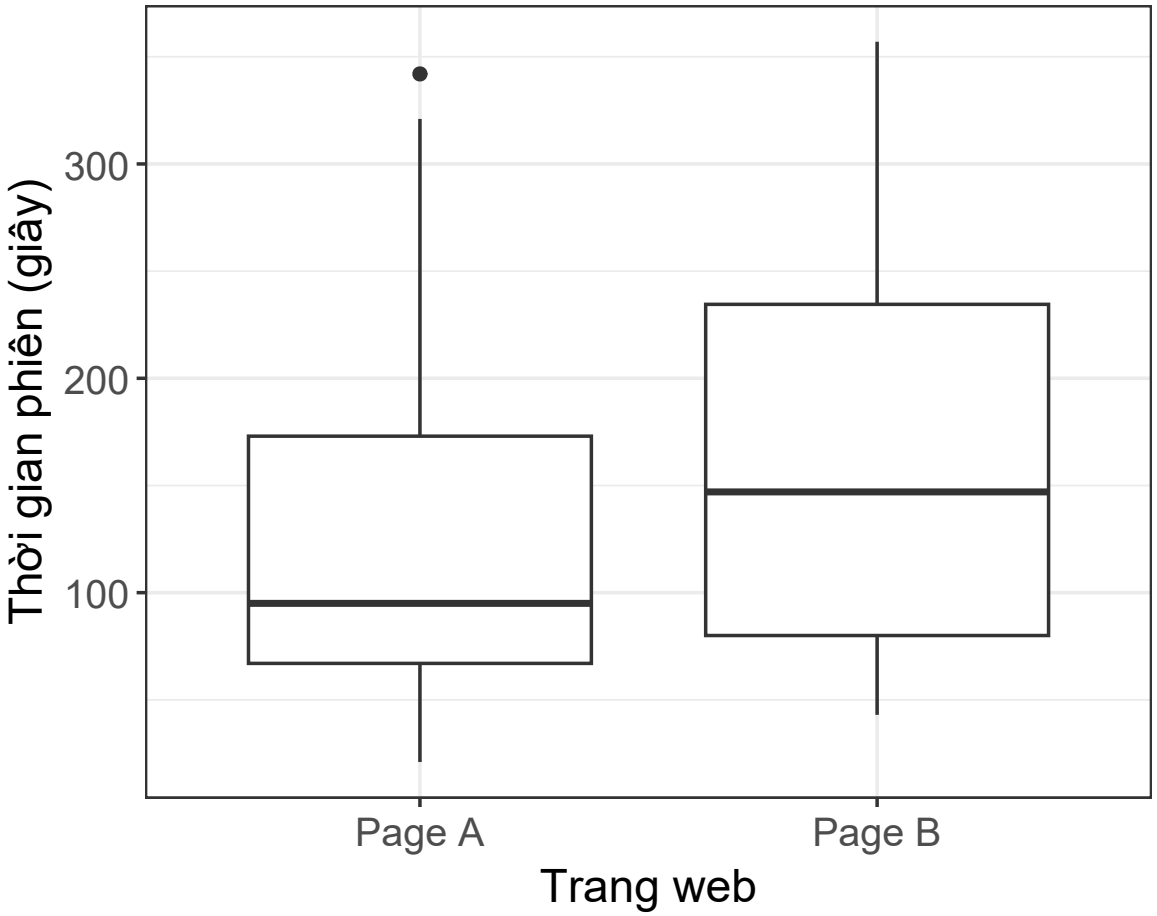
Cách tiếp cận

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: Tăng tính tương tác tới trang web



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

Ví dụ: Tăng tính tương tác tới trang web

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 như sau:

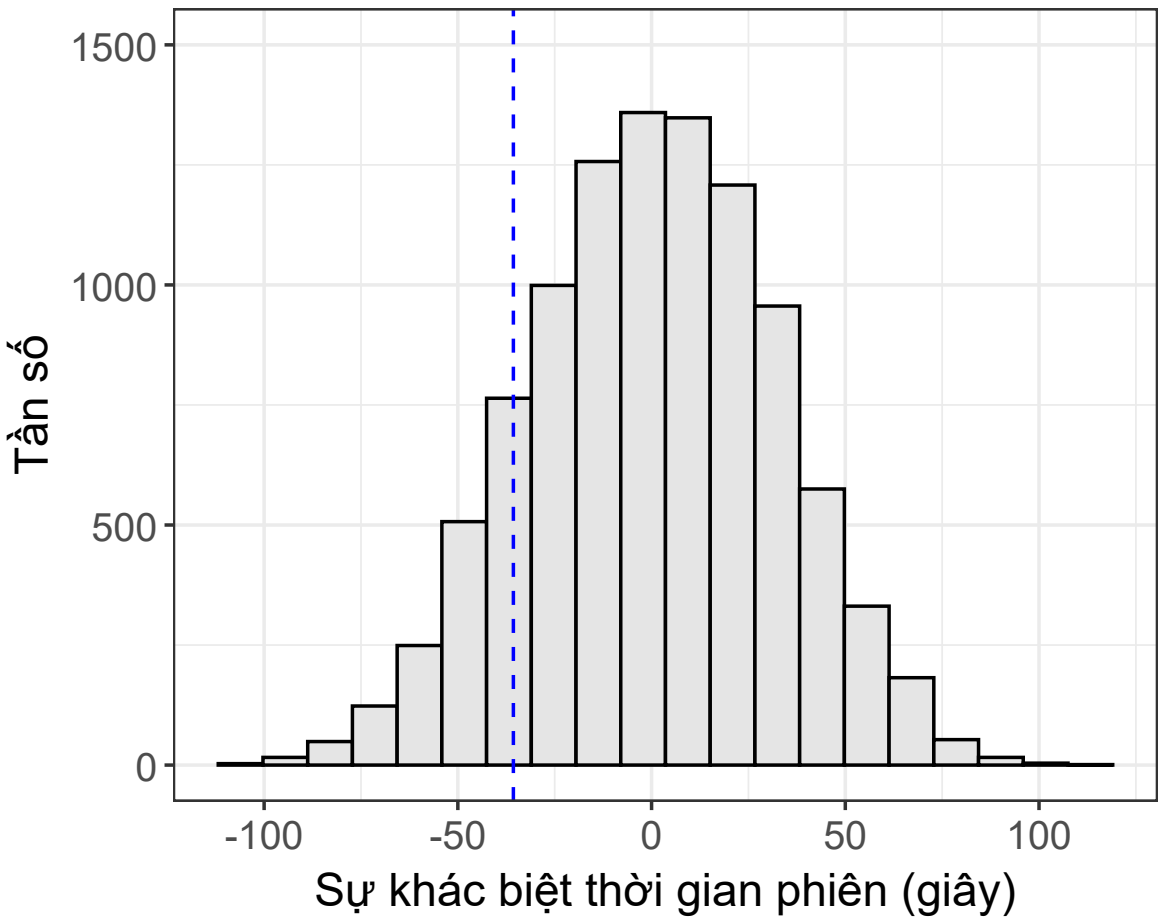
$$\begin{cases} H_0: & \mu_A \geq \mu_B, \\ H_1: & \mu_A < \mu_B. \end{cases}$$

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;
- $\Delta = \overline{y}_A - \overline{y}_B = -35.67$.

Ta thực hiện kiểm định hoán vị với $R = 10000$ lần.

Ví dụ: Tăng tính tương tác tới trang web



Ta tính được

$$p\text{-value} = \frac{1}{10000}\sum_{j=1}^{10000} \mathrm{I}(t_{\mathrm{diff},j} < -35.67) = 0.1333$$

Ví dụ: Tăng tính tương tác tới trang web

Chọn mức ý nghĩa $\alpha = 0.05$.

$\Rightarrow$ p-value $= 0.1333 > \alpha$

$\Rightarrow$ ta không đủ cơ sở để bác bỏ Giả thuyết không H_0.

$\Rightarrow$ 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.

Khoảng tin cậy cho p-value

Quá trình permutation test thực sự yêu cầu $n!$ lần xáo trộn mẫu

$\Rightarrow$ p-value thu được sẽ là giá trị **thực sự**.

Tuy nhiên, ta chỉ thực hiện R lần xáo trộn (bé hơn $n!$)

$\Rightarrow$ p-value thu được chỉ là 1 trong số rất nhiều giá trị có thể.

Lần chạy	R	p-value
1	10000	0.1333
2	10000	0.1291
3	10000	0.1349
4	10000	0.1380
5	10000	0.1306
6	10000	0.1419

Khoảng tin cậy cho p-value

Câu hỏi: Vậy giá trị p-value thực sự là bao nhiêu?

Trả lời: Rất khó (không thể) biết được chính xác.

Tuy nhiên, ta có thể xác định khoảng tin cậy 99% cho p-value

$$\left(\widehat{p} - z_{0.995}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{R}}, \widehat{p} + z_{0.995}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{R}}\right),$$

trong đó $z_{0.995}$ là phân vị thứ 99% của phân phối chuẩn $\mathcal{N}(0,1)$.

Ví dụ: xét lại ví dụ kiểm định thời gian tương tác của hai trang web A và B.

Khoảng tin cậy 99% cho p-value là

$$\left(0.1333 - 2.5758\sqrt{\frac{0.1333 \times (1-0.1333)}{10000}}, 0.1333 + 2.5758\sqrt{\frac{0.1333 \times (1-0.1333)}{10000}}\right)$$

$$= (0.1245, 0.1421).$$

$\Rightarrow$ gần như chắc chắn rằng p-value của permutation test là lớn hơn $\alpha = 0.05$.

1 Tổng quan

2 Hai nhóm độc lập

3 Hai nhóm ghép cặp

4 Bootstrap test một mẫu

5 Bootstrap test cho 2 mẫu

Bài toán

Ví dụ: Xét mẫu ghép cặp gồm cân nặng trước và sau khi thực hiện 1 chế độ ăn kiêng của 12 người.

Người thứ	Cân nặng trước	Cân nặng sau
1	157	150
2	185	181
3	120	121
4	212	206
5	230	215
6	165	169
7	207	210
8	251	232
9	196	188
10	140	138
11	137	145
12	172	172

Câu hỏi: Liệu rằng chế độ ăn kiêng có làm thay đổi cân nặng?

Bài toán

Vì cân nặng tới cùng từ 1 người, ta tính được sự thay đổi cân nặng bằng phép trừ (trước − sau):

Người thứ	Cân nặng trước	Cân nặng sau	thay đổi (d)
1	157	150	7
2	185	181	4
3	120	121	−1
4	212	206	6
5	230	215	15
6	165	169	−4
7	207	210	−3
8	251	232	19
9	196	188	8
10	140	138	2
11	137	145	−8
12	172	172	0

- nếu trung bình mức thay đổi cân nặng bằng 0 ⇒ ăn kiêng không hiệu quả;
- nếu trung bình mức thay đổi cân nặng khác 0 ⇒ ăn kiêng có tác động tới cân nặng.

Paired Permutation Test

Xét hai nhóm dữ liệu ghép cặp

- $(Y_{A,1}, Y_{A,2}, \ldots, Y_{A,n})$ độc lập cùng phân phối F_{Y_A};
- $(Y_{B,1}, Y_{B,2}, \ldots, Y_{B,n})$ độc lập cùng phân phối F_{Y_B}.

Xét bài toán kiểm định

$$\begin{cases} H_0: & \text{không có hiệu quả} \\ H_1: & \text{có hiệu quả} \end{cases}$$

Giả thuyết không H_0 tương đương với $F_{Y_A} = F_{Y_B} = F_Y$, khi đó, việc hoán đổi nhãn của hai nhóm dữ liệu cũng sẽ không thay đổi phân phối của chúng.

Paired Permutation Test

Dữ liệu là ghép cặp $\Rightarrow$ không thể hoán đổi chéo các đối tượng với nhau!

Trong dữ liệu ghép cặp:

- mỗi cặp $(Y_{A,i}, Y_{B,i})$ là một đơn vị độc lập,
- dưới giả thuyết H_0, việc đổi chỗ $Y_{A,i}$ cho $Y_{B,i}$ trong từng cặp là có khả năng như nhau

Do đó $(Y_{A,i}, Y_{B,i})$ và $(Y_{B,i}, Y_{A,i})$ có xác suất như nhau dưới H_0.

Việc hoán vị này cho ta: $D_i = Y_{A,i} - Y_{B,i}$ hoặc $-D_i$.

$\Rightarrow$ Hoán vị cho dữ liệu ghép cặp $\Leftrightarrow$ tạo mẫu ngẫu nhiên $D_1^*, \ldots, D_n^*$:

$$D_i^* = \varepsilon_i D_i,$$

trong đó, $\varepsilon_i \in \{-1, 1\}$ được chọn ngẫu nhiên.

Random sign permutation hoặc **Sign-flipping test**.

Paired Permutation Test

Sử dụng sự khác biệt trung bình làm thống kê kiểm định.

$\Rightarrow$ cặp giả thuyết được viết lại thành

$$\begin{cases} H_0: & \mu_D = 0, \\ H_1: & \mu_D \neq 0, \end{cases}$$

với $\mu_D = \mu_A - \mu_B$.

Khi này, giá trị thống kê kiểm định trên mẫu sẽ là $t_{\text{diff}} = \overline{Y}_A - \overline{Y}_B$.

Nếu giả thuyết H_0 là đúng thì sẽ rất bất ngờ nếu ta quan sát thấy t_{diff} là lớn hơn hoặc nhỏ hơn hầu hết các giá trị $t_{\text{diff},j}$ (với $j = 1, \ldots, 2^n$, 2^n là tổng số lượng hoán vị).

Khi này, ta định nghĩa được p-value là

$$p\text{-value} = \frac{1}{2^n} \sum_{j=1}^{2^n} \mathrm{I}\left(|t_{\text{diff},j}| \geq |t_{\text{diff}}|\right).$$

Nếu p-value $\leq \alpha \Rightarrow$ giả thuyết gốc H_0 là không tương thích với dữ liệu, hay việc hoán đổi thứ tự trong 1 cặp là **không** ngẫu nhiên

$\Rightarrow$ có sự khác biệt về giá trị của hai nhóm.

Paired Permutation Test

Quy trình kiểm định hoán vị cho dữ liệu ghép cặp:

Bước 1 Với mỗi cặp quan sát $(Y_{A,i}, Y_{B,i})$, tính hiệu số $D_i = Y_{A,i} - Y_{B,i}$ với $i = 1, \ldots, n$.

Bước 2 Tính thống kê kiểm định từ dữ liệu gốc:

$$t_{\text{diff}} = \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i.$$

Bước 3 Tạo ngẫu nhiên một vector dấu $(\varepsilon_1, \ldots, \varepsilon_n)$, trong đó $\varepsilon_i \in \{-1, 1\}$, với xác suất bằng nhau (tức là 1/2).

Bước 4 Tạo dữ liệu hoán vị: $D_{i,\pi} = \varepsilon_i D_i$, $i = 1, \ldots, n$.

Bước 5 Tính thống kê kiểm định từ dữ liệu hoán vị:

$$t_{\text{diff},\pi} = \frac{1}{n}\sum_{i=1}^{n} D_{i,\pi}.$$

Bước 6 Lặp lại bước 3–5 trong 2^n lần (tương ứng với tổng số cấu hình đổi dấu có thể xảy ra).

Bước 7 Tính p-value theo công thức và so sánh với mức ý nghĩa α.

Paired Permutation Test

Tuy nhiên, 2^n sẽ tăng rất nhanh khi số lượng cặp quan sát lớn. Chẳng hạn:

- $2^{10} = 1024$,
- $2^{15} = 32768$,
- $2^{20} = 1048576$,
- $2^{30} \approx 10^9$.

$\Rightarrow$ chỉ lấy ngẫu nhiên R lần đổi dấu thay vì xét toàn bộ 2^n khả năng.

Vì vậy, trong áp dụng thực tế, bước 5 và 6 sẽ được thay thế bởi:

*Bước 6** Lặp lại bước 3 – 5 trong R lần (tối thiểu 1,000 lần).

*Bước 7** Tính p-value theo công thức

$$p\text{-value} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}\left(|t_{\mathrm{diff},j}| \geq |t_{\mathrm{diff}}|\right),$$

và so sánh với mức ý nghĩa α.

Paired Permutation Test

Trong trường hợp ta muốn kiểm định các cặp giả thuyết một phía

$$\begin{cases} H_0 : \mu_A \leq \mu_B \\ H_1 : \mu_A > \mu_B \end{cases}, \qquad \begin{cases} H_0 : \mu_A \geq \mu_B \\ H_1 : \mu_A < \mu_B \end{cases}$$

khi đó, p-value được xác định bởi:

- cặp giả thuyết kiểm định một phía bên phải

$$p\text{-value} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}\left(t_{\mathrm{diff},j} \geq t_{\mathrm{diff}}\right);$$

- cặp giả thuyết kiểm định một phía bên trái

$$p\text{-value} = \frac{1}{R}\sum_{j=1}^{R} \mathrm{I}\left(t_{\mathrm{diff},j} \leq t_{\mathrm{diff}}\right).$$

Paired Permutation Test

Ta có thể xác định khoảng tin cậy 99% cho p-value

$$\left(\widehat{p} - z_{0.995}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{R}}, \widehat{p} + z_{0.995}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{R}}\right),$$

trong đó $z_{0.995}$ là phân vị thứ 99% của phân phối chuẩn $\mathcal{N}(0,1)$.

Ví dụ: Hiệu quả ăn kiêng

Ta cần kiểm tra xem cân nặng trung bình của người sau khi ăn kiêng (B) có thực sự nhỏ hơn so với trước khi ăn kiêng (A), hay không.

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

$$\begin{cases} H_0: & \mu_D \leq 0 \\ H_1: & \mu_D > 0, \end{cases}$$

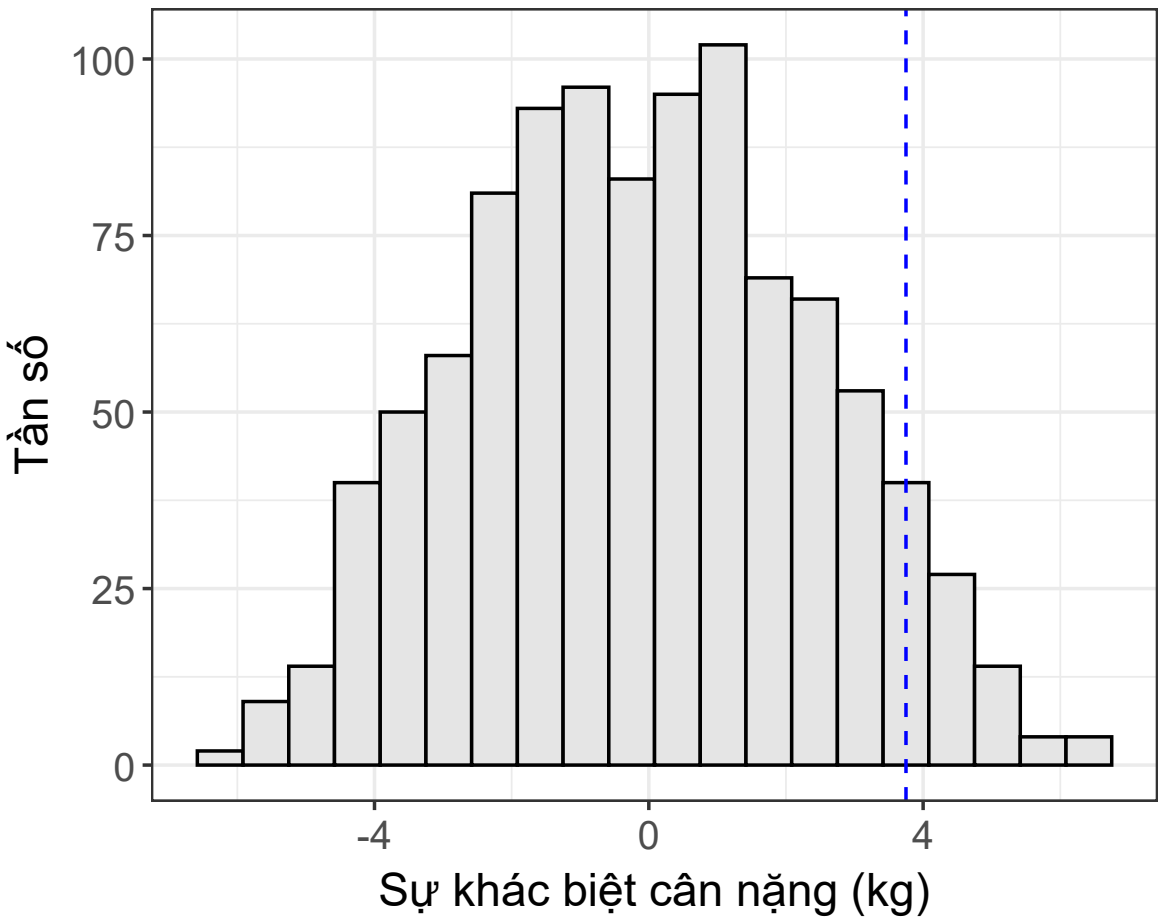
với $\mu_D = \mu_A - \mu_B$.

Từ dữ liệu ta có một số thông tin

- cỡ mẫu $n = 12$ (cần $2^{12} = 4096$ hoán vị);
- trung bình chênh lệch cân nặng là $\overline{D} = \overline{Y}_A - \overline{Y}_B = 3.75$ kg;
- $t_{\text{diff}} = \overline{Y}_A - \overline{Y}_B = 3.75$.

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

Ví dụ: Hiệu quả ăn kiêng



Ta tính được

$$p\text{-value} = \frac{1}{1000}\sum_{j=1}^{1000} \mathrm{I}(t_{\mathrm{diff},j} \geq 3.57) = 0.077$$

Ví dụ: Hiệu quả ăn kiêng

Khoảng tin cậy 99% cho p-value là

$$\left(0.077 - 2.5758\sqrt{\frac{0.077 \times (1 - 0.077)}{1000}}, 0.077 + 2.5758\sqrt{\frac{0.077 \times (1 - 0.077)}{1000}}\right)$$

$$= (0.0553, 0.0987).$$

$\Rightarrow$ gần như chắc chắn rằng p-value của permutation test là bé hơn $\alpha = 0.1$.

$\Rightarrow$ ta đủ cơ sở để bác bỏ Giả thuyết không H_0.

$\Rightarrow$ Sự giảm cân nặng trung bình sau khi ăn kiêng so với trước là có **ý nghĩa thống kê** (**statistically significant**).

Kết luận: Ăn kiêng có tác động tích cực tới cân nặng, với mức ý nghĩa 10%.

1 Tổng quan

2 Hai nhóm độc lập

3 Hai nhóm ghép cặp

4 Bootstrap test một mẫu

5 Bootstrap test cho 2 mẫu

Trung bình một mẫu

Cho mẫu $Y_1, Y_2, \ldots, Y_n$ độc lập cùng phân phối F có kỳ vọng hữu hạn μ.

Ta mong muốn kiểm định giả thuyết

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

Để kiểm định giả thuyết này ta có thể dùng một trong hai thống kê kiểm định:

- $T(Y_1, Y_2, \ldots, Y_n) = \overline{Y} - \mu_0$; hoặc
- $T(Y_1, Y_2, \ldots, Y_n) = \sqrt{n}\dfrac{\overline{Y} - \mu_0}{\widehat{\sigma}}$.

Giả sử rằng $F \equiv F_0$ đã biết, khi đó,

$$p\text{-value} = \Pr_{F_0}\left(\left|\overline{Y} - \mu_0\right| \geq |t_{\text{obs}}|\right)$$

hoặc

$$p\text{-value} = \Pr_{F_0}\left(\sqrt{n}\frac{\left|\overline{Y} - \mu_0\right|}{\widehat{\sigma}} \geq |t_{\text{obs}}|\right).$$

Trung bình một mẫu

Đặt $\widehat{F}_n$ là phân phối thực nghiệm của dữ liệu.

Khi đó, với mẫu $Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*$ được tạo từ $\widehat{F}_n$, ta xác định được

- $T(Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*) = \overline{Y}_j^* - \overline{Y}$; hoặc
- $T(Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*) = \sqrt{n}\dfrac{\overline{Y}_j^* - \overline{Y}}{\widehat{\sigma}_j^*}$.

p-value được xấp xỉ bởi (áp dụng công thức slide 13, thế μ_0 bằng $\mathbb{E}(Y^*) = \overline{Y}$)

$$p\text{-value} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\left|\overline{Y}_j^* - \overline{Y}\right| \geq |t_{\text{obs}}|\right)$$

hoặc

$$p\text{-value} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\sqrt{n}\frac{\left|\overline{Y}_j^* - \overline{Y}\right|}{\widehat{\sigma}_j^*} \geq |t_{\text{obs}}|\right),$$

với

- $t = T(y_1, y_2, \ldots, y_n) = \overline{y} - \mu_0$; hoặc
- $t = T(y_1, y_2, \ldots, y_n) = \sqrt{n}\dfrac{\overline{y} - \mu_0}{\widehat{\sigma}}$.

Trung bình một mẫu

Tuy nhiên, p-value được xấp xỉ chỉ dùng được trong kiểm định nếu mẫu $Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*$ được tạo dưới giả định H_0.

Trung bình một mẫu

Tuy nhiên, p-value được xấp xỉ chỉ dùng được trong kiểm định nếu mẫu $Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*$ được tạo dưới giả định H_0.

Câu hỏi: tạo $Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*$ từ $\widehat{F}_n$ có tương đương với tạo dưới giả định H_0?

Trung bình một mẫu

Tuy nhiên, p-value được xấp xỉ chỉ dùng được trong kiểm định nếu mẫu $Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*$ được tạo dưới giả định H_0.

Câu hỏi: tạo $Y_{1j}^*, Y_{2j}^*, \ldots, Y_{nj}^*$ từ $\widehat{F}_n$ có tương đương với tạo dưới giả định H_0?

Nhắc lại rằng,

- dưới giả định H_0: $\mathbb{E}(Y) = \mu_0$;
- Y^* được tạo từ $\widehat{F}_n$: $\mathbb{E}_{\widehat{F}_n}(Y^*) = \overline{Y} \neq \mu_0$.

Do đó, để đảm bảo mẫu được tạo dưới H_0, thì $\mathbb{E}_{\widehat{F}_n}(Z^*) = \mu_0$. Khi đó, với

$$Z^* = Y^* - \overline{Y} + \mu_0$$

thì

- $\mathbb{E}_{\widehat{F}_n}(Z^*) = \mu_0$;
- $\mathbb{V}\mathrm{ar}_{\widehat{F}_n}(Z^*) = \mathbb{V}\mathrm{ar}_{\widehat{F}_n}(Y^*)$.

Thống kê kiểm định sẽ là

- $T(Z_{1j}^*, Z_{2j}^*, \ldots, Z_{nj}^*) = \overline{Z}_j^* - \mu_0$; hoặc
- $T(Z_{1j}^*, Z_{2j}^*, \ldots, Z_{nj}^*) = \sqrt{n}\dfrac{\overline{Z}_j^* - \mu_0}{\widehat{\sigma}_j^*}$.

Trung bình một mẫu

Khi này, p-value được xấp xỉ bởi (áp dụng công thức slide 13)

$$p\text{-value} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\left|\overline{Z}_j^* - \mu_0\right| \geq \left|t_{\text{obs}}\right|\right)$$

hoặc

$$p\text{-value} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\sqrt{n}\frac{\left|\overline{Z}_j^* - \mu_0\right|}{\widehat{\sigma}_j^*} \geq \left|t_{\text{obs}}\right|\right),$$

với

- $t_{\text{obs}} = T(y_1, y_2, \ldots, y_n) = \overline{y} - \mu_0$; hoặc
- $t_{\text{obs}} = T(y_1, y_2, \ldots, y_n) = \sqrt{n}\dfrac{\overline{y} - \mu_0}{\widehat{\sigma}}$.

Lúc này, xấp xỉ p-value tương ứng của $Z_{1j}^*, Z_{2j}^*, \ldots, Z_{nj}^*$ có thể dùng được trong kiểm định.

Trung bình một mẫu

Mặt khác, ta chứng minh được rằng

$$\overline{Z}_j - \mu_0 = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}^* - \mu_0 = \overline{Y}_j^* - \overline{Y}.$$

Điều này dẫn tới

$$p\text{-value} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\left|\overline{Z}_j^* - \mu_0\right| \geq |t_{\text{obs}}|\right) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\left|\overline{Y}_j^* - \overline{Y}\right| \geq |t_{\text{obs}}|\right)$$

hoặc

$$p\text{-value} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\sqrt{n}\frac{\left|\overline{Z}_j^* - \mu_0\right|}{\widehat{\sigma}_j^*} \geq |t_{\text{obs}}|\right) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{I}\left(\sqrt{n}\frac{\left|\overline{Y}_j^* - \overline{Y}\right|}{\widehat{\sigma}_j^*} \geq |t_{\text{obs}}|\right).$$

Do đó, quá trình tạo Y^* từ $\widehat{F}_n$ thực sự kết hợp tính chất H_0.

Kiểm định bootstrap cho trung bình

Cho mẫu $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} F$, có trung bình μ và phương sai σ^2 chưa biết.

Xét bài toán kiểm định cho trung bình

$$H_0 : \mu = \mu_0 \qquad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

Đặt $\widehat{F}_n$ là ước lượng thực nghiệm của F.

Từ dữ liệu ta tính được

- trung bình mẫu $\overline{Y}$;
- phương sai mẫu $\widehat{\sigma}^2$.

Kiểm định bootstrap cho trung bình

Kiểm định bootstrap cho trung bình

Quá trình kiểm định bootstrap được mô tả như sau:

1. Ta tạo một mẫu ngẫu nhiên $Y_1^*, Y_2^*, \ldots, Y_n^*$ (cùng với cỡ của dữ liệu gốc) từ dữ liệu gốc, có lặp lại (with replacement).
2. Tính ước lượng $\overline{Y}^*$ (và $\widehat{\sigma}^*$) từ dữ liệu vừa tạo.
3. Tính giá trị thống kê

$$t^* = \overline{Y}^* - \overline{Y}, \qquad \text{hoặc} \qquad t^* = \sqrt{n}\frac{\overline{Y}^* - \overline{Y}}{\widehat{\sigma}^*},$$

4. Lặp lại bước 1 và 2 trong B lần (ít nhất 2000 lần), và lưu kết quả lại.
5. Tính p-value theo công thức

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\left|t_j^*\right| \geq |t_{\text{obs}}|\right),$$

với

$$t_{\text{obs}} = \overline{Y} - \mu_0, \qquad \text{hoặc} \qquad t_{\text{obs}} = \sqrt{n}\frac{\overline{Y} - \mu_0}{\widehat{\sigma}}.$$

Kiểm định bootstrap cho trung bình - Ví dụ 1 - Phân phối chuẩn

Ví dụ 1: Ta xét mô phỏng kiểm định cho trung bình của một mẫu:

$$Y_1, \ldots, Y_n \overset{i.i.d}{\sim} \mathcal{N}(0,1)$$

Hai trường hợp giả thuyết

- trường hợp 1:

$$H_0 : \mu = 0 \qquad \text{vs.} \quad H_1 : \mu \neq 0$$

  tức là H_0 đúng;

- trường hợp 2:

$$H_0 : \mu = 1 \qquad \text{vs.} \quad H_1 : \mu \neq 1$$

  tức là H_0 sai.

Trong cả hai trường hợp, ta thiết lập

- $n = 10, 20, 30$;
- $\alpha = 0.1, 0.05, 0.01$;
- $B = 2000$ lần (bootstrap);
- lặp lại $M = 5000$ lần (Monte Carlo).

Kiểm định bootstrap cho trung bình - Ví dụ 1 - Phân phối chuẩn

Trong cả trường hợp, ta áp dụng 3 phương pháp kiểm định:

- t-test (giả định phân phối chuẩn);
- bootstrap với thống kê kiểm định $t^* = \overline{Y}^* - \overline{Y}$ (bootstrap-diff);
- bootstrap với thống kê kiểm định $t^* = \sqrt{n}\dfrac{\overline{Y}^* - \overline{Y}}{\widehat{\sigma}^*}$ (bootstrap-pivot);

Đối với trường hợp 1, H_0 đúng, ta tính xác suất sai lầm loại I thực nghiệm.

		Phương pháp		
	Cỡ mẫu	t-test	bootstrap-diff	bootstrap-pivot
$\alpha = 0.1$	$n = 10$	0.1026	0.1564	0.0922
	$n = 20$	0.0952	0.1242	0.0936
	$n = 30$	0.0996	0.1186	0.0990
$\alpha = 0.05$	$n = 10$	0.0522	0.0998	0.0428
	$n = 20$	0.0526	0.0698	0.0498
	$n = 30$	0.0518	0.0632	0.0504
$\alpha = 0.01$	$n = 10$	0.0116	0.0418	0.0088
	$n = 20$	0.0088	0.0226	0.0082
	$n = 30$	0.0122	0.0224	0.0116

Kiểm định bootstrap cho trung bình - Ví dụ 1 - Phân phối chuẩn

Đối với trường hợp 2, H_0 sai, ta tính độ mạnh thực nghiệm.

		Phương pháp		
	Cỡ mẫu	t-test	bootstrap-diff	bootstrap-pivot
$\alpha = 0.1$	$n = 10$	0.8982	0.9396	0.8808
	$n = 20$	0.9964	0.9970	0.9966
	$n = 30$	1.0000	1.0000	1.0000
$\alpha = 0.05$	$n = 10$	0.8068	0.8954	0.7588
	$n = 20$	0.9884	0.9938	0.9864
	$n = 30$	1.0000	1.0000	1.0000
$\alpha = 0.01$	$n = 10$	0.5090	0.7696	0.3852
	$n = 20$	0.9324	0.9666	0.9140
	$n = 30$	0.9970	0.9988	0.9968

Kiểm định bootstrap cho trung bình - Ví dụ 2 - Phân phối mũ

Ví dụ 2: Ta xét mô phỏng kiểm định cho trung bình của một mẫu:

$$Y_1, \ldots, Y_n \overset{i.i.d}{\sim} \mathrm{Exp}(1)$$

Hai trường hợp giả thuyết

- trường hợp 1:

$$H_0 : \mu = 1 \qquad \text{vs.} \quad H_1 : \mu \neq 1$$

  tức là H_0 đúng;

- trường hợp 2:

$$H_0 : \mu = 2 \qquad \text{vs.} \quad H_1 : \mu \neq 2$$

  tức là H_0 sai.

Trong cả hai trường hợp, ta thiết lập

- $n = 10, 20, 30$;
- $\alpha = 0.1, 0.05, 0.01$;
- $B = 2000$ lần (bootstrap);
- lặp lại $M = 5000$ lần (Monte Carlo).

Kiểm định bootstrap cho trung bình - Ví dụ 2 - Phân phối mũ

Trong cả trường hợp, ta áp dụng 3 phương pháp kiểm định:

- t-test (giả định phân phối chuẩn);
- bootstrap với thống kê kiểm định $t^* = \overline{Y}^* - \overline{Y}$ (bootstrap-diff);
- bootstrap với thống kê kiểm định $t^* = \sqrt{n}\dfrac{\overline{Y}^* - \overline{Y}}{\widehat{\sigma}^*}$ (bootstrap-pivot);

Đối với trường hợp 1, H_0 đúng, ta tính xác suất sai lầm loại I thực nghiệm.

		Phương pháp		
	Cỡ mẫu	t-test	bootstrap-diff	bootstrap-pivot
$\alpha = 0.1$	$n = 10$	0.1464	0.1970	0.1208
	$n = 20$	0.1300	0.1586	0.1152
	$n = 30$	0.1242	0.1438	0.1098
$\alpha = 0.05$	$n = 10$	0.1002	0.1486	0.0710
	$n = 20$	0.0834	0.1070	0.0654
	$n = 30$	0.0736	0.0898	0.0610
$\alpha = 0.01$	$n = 10$	0.0442	0.0864	0.0218
	$n = 20$	0.0374	0.0530	0.0194
	$n = 30$	0.0286	0.0402	0.0150

Kiểm định bootstrap cho trung bình - Ví dụ 2 - Phân phối mũ

Đối với trường hợp 2, H_0 sai, ta tính độ mạnh thực nghiệm.

		Phương pháp		
	Cỡ mẫu	t-test	bootstrap-diff	bootstrap-pivot
$\alpha = 0.1$	$n = 10$	0.8364	0.8842	0.7400
	$n = 20$	0.9674	0.9760	0.9372
	$n = 30$	0.9954	0.9968	0.9894
$\alpha = 0.05$	$n = 10$	0.7652	0.8386	0.6120
	$n = 20$	0.9360	0.9562	0.8656
	$n = 30$	0.9870	0.9906	0.9666
$\alpha = 0.01$	$n = 10$	0.5820	0.7262	0.3490
	$n = 20$	0.8420	0.8830	0.6408
	$n = 30$	0.9552	0.9644	0.8510

Kiểm định bootstrap cho trung bình - kiểm định một phía

Ta có thể áp dụng quy trình kiểm định bootstrap ở trên cho bài toán kiểm định một phía

$$\begin{cases} H_0: & \mu \geq \mu_0 \\ H_1: & \mu < \mu_0 \end{cases} \quad \text{hoặc} \quad \begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

Quy trình bootstrap sẽ là tương tự như trường hợp kiểm định hai phía.

p-value sẽ được tính bởi

- kiểm định một phía bên trái ($<$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(t_j^* < t_{\text{obs}}\right),$$

- kiểm định một phía bên phải ($>$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(t_j^* > t_{\text{obs}}\right).$$

1 Tổng quan

2 Hai nhóm độc lập

3 Hai nhóm ghép cặp

4 Bootstrap test một mẫu

5 Bootstrap test cho 2 mẫu

Bài toán kiểm định hai mẫu - A/B testing

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 kiểm định hai mẫu - A/B testing

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.

Trung bình hai mẫu - đồng phương sai

Cho hai mẫu ngẫu nhiên

- $X_1, \ldots, X_{n_1} \overset{i.i.d}{\sim} F$;
- $Y_1, \ldots, Y_{n_2} \overset{i.i.d}{\sim} G$,

với $\mathbb{E}(X) = \mu_X$ và $\mathbb{E}(Y) = \mu_Y$.

Ta giả sử rằng

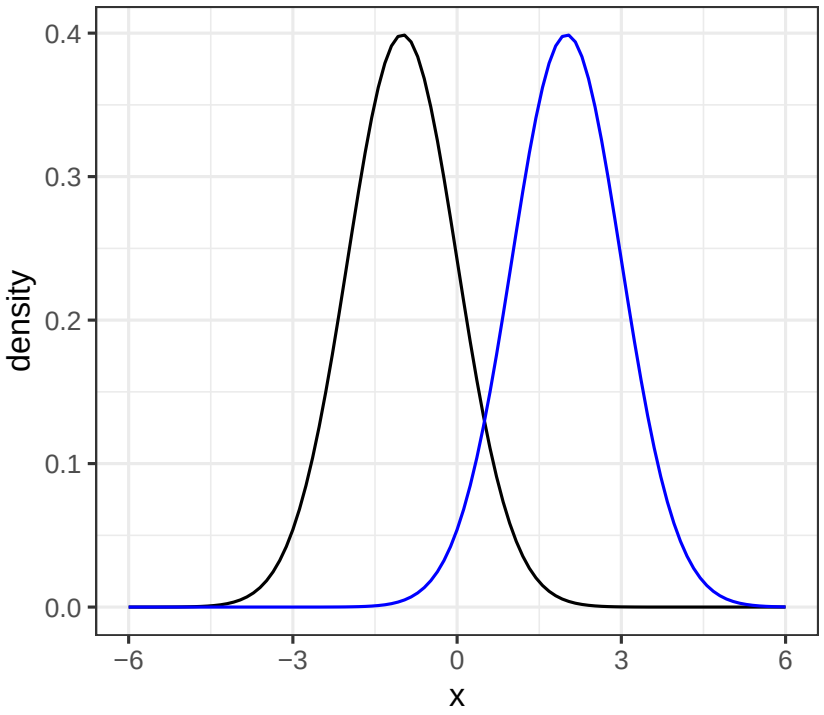
$$F(t) = H(t - \mu_X) \qquad \text{và} \qquad G(t) = H(t - \mu_Y),$$

với mọi $t \in \mathbb{R}$, và H là một phân phối chưa biết, có phương sai σ^2.

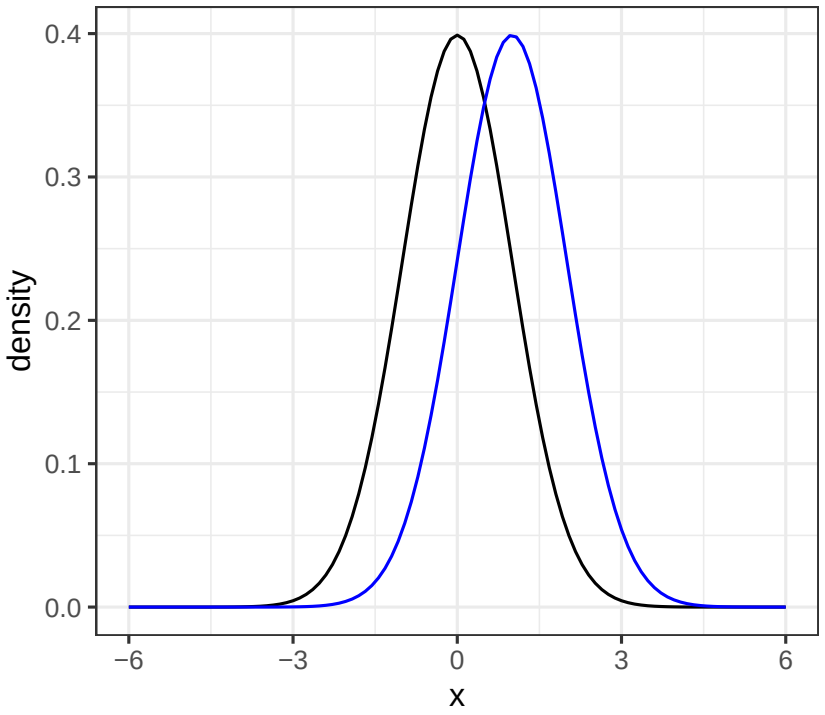
Giả định này đảm bảo rằng

- X và Y có cùng phương sai;
- $F = G$ tương đương với $\mu_X = \mu_Y$.

Trung bình hai mẫu - đồng phương sai



(a) $\mu_X = -1, \mu_Y = 2$



(b) $\mu_X = 0, \mu_Y = 1$

Trung bình hai mẫu - đồng phương sai

Ta quan tâm tới kiểm định giả thuyết

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_1 : \mu_X \neq \mu_Y.$$

Theo lý thuyết kiểm định:

- $S_p = \dfrac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}$,
- thống kê kiểm định

$$T_{n_1,n_2} = \frac{(\overline{X} - \overline{Y})}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2},$$

nếu F và G là phân phối chuẩn và H_0 đúng.

Mặt khác, nếu $n_1/n_2 \to \lambda$ và $n_1 \to \infty$, theo luật số lớn dạng mạnh và định lý giới hạn trung tâm, ta chứng minh được

$$T_{n_1,n_2} \to \mathcal{N}\left(0, \frac{1}{1+\lambda} + \frac{1}{1+1/\lambda}\right) \equiv \mathcal{N}(0,1).$$

Trung bình hai mẫu - đồng phương sai

p-value được xác định bởi

$$p\text{-value} = \Pr_{F_0}\left(\frac{|\overline{X}-\overline{Y}|}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > |t_{obs}|\right),$$

với phân phối F_0 được xác định là:

- t-student nếu F, G là phân phối chuẩn;
- phân phối chuẩn nếu n_1, n_2 lớn.

Trong thực tế,

- dữ liệu thường không ủng hộ giả thuyết phân phối chuẩn cho F và G;
- tỷ lệ cỡ mẫu n_1/n_2 khó đảm bảo đặc biệt là cỡ mẫu lớn.

Do đó, việc áp dụng

- kiểm định t-test hai mẫu;
- xấp xỉ chuẩn theo định lý giới hạn trung tâm;

không cho kết quả tốt.

Trung bình hai mẫu - đồng phương sai

Ta áp dụng bootstrap để tạo mẫu phỏng theo giả thuyết H_0:

- X và Y có cùng phân phối;
- $\sigma_X^2 = \sigma_Y^2 = \sigma^2$;

$\Rightarrow$ ta có thể ghép hai mẫu này thành một (mẫu gộp - pooled sample):

$$Z_1, \ldots, Z_{n_1+n_2} = X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}$$

Kiểm định bootstrap cho hai trung bình, đồng phương sai

1. Ta tạo một mẫu ngẫu nhiên $Z_1^*, Z_2^*, \ldots, Z_{n_1+n_2}^*$ (cùng với cỡ của dữ liệu gốc) từ dữ liệu gốc, có lặp lại (with replacement).
2. Từ mẫu bootstrap, ấn định

   n_1 giá trị đầu là mẫu bootstrap của X: $X_1^*, X_2^*, \ldots, X_{n_1}^*$

   $n - n_1$ giá trị tiếp theo là mẫu bootstrap của Y: $Y_1^*, Y_2^*, \ldots, Y_{n_2}^*$.
3. Tính ước lượng $\overline{X}^*$, S_X^{*2}, $\overline{Y}^*$ và S_Y^{*2} từ dữ liệu vừa tạo.

Trung bình hai mẫu - đồng phương sai

4. Tính giá trị thống kê

$$t^* = \overline{X}^* - \overline{Y}^*, \qquad \text{hoặc} \qquad t^* = \frac{(\overline{X}^* - \overline{Y}^*)}{S_p^* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

5. Lặp lại bước 1, 2, 3 và 4 trong B lần (ít nhất 2000 lần), và lưu kết quả lại.
6. Tính p-value theo công thức

$$p\text{-value} = \frac{1}{B} \sum_{j=1}^{B} \mathrm{I}\left(\left|t_j^*\right| \geq |t_{\text{obs}}|\right),$$

với

$$t_{\text{obs}} = \overline{X} - \overline{Y}, \qquad \text{hoặc} \qquad t_{\text{obs}} = \frac{(\overline{X} - \overline{Y})}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Trung bình hai mẫu - đồng phương sai

Ví dụ: Một nhóm nghiên cứu đã thu thập cá trong hồ Whitewater và hồ Wirth tại Minnesota, Mỹ, và đo hàm lượng thủy ngân trong mỗi con cá.

Nhóm muốn tìm hiểu xem liệu hàm lượng thủy ngân trong cá ở 2 hồ có bằng nhau hay không?

	n	mean	sd
Whitewater	9	0.2962	0.0376
Wirth	9	0.2572	0.0248

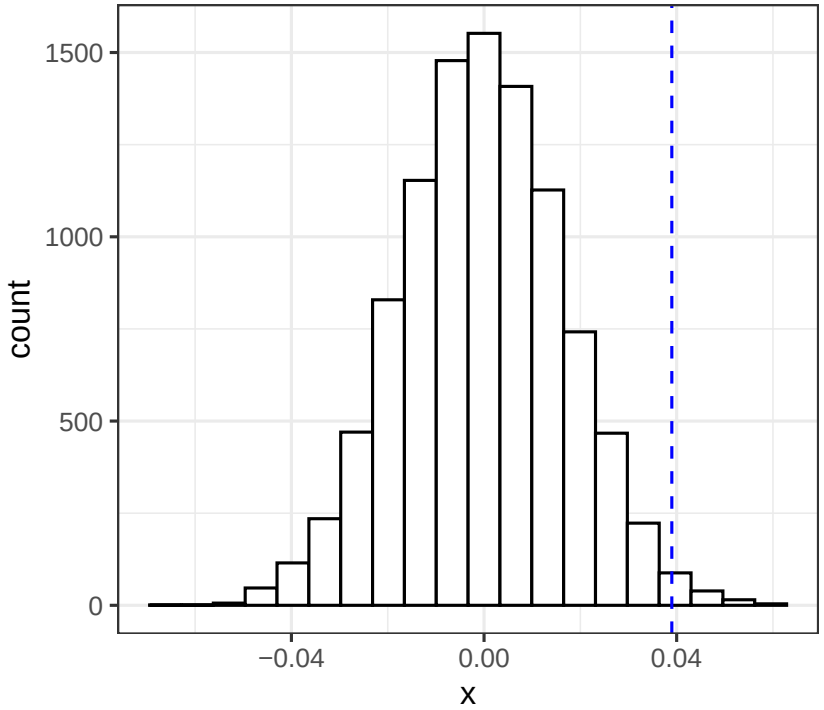
Nhận thấy rằng, tỷ số của hai độ lệch chuẩn là 1.5155, nằm trong (0.5, 2).

Do đó, ta có thể giả định rằng phương sai là đồng nhất giữa hai nhóm.

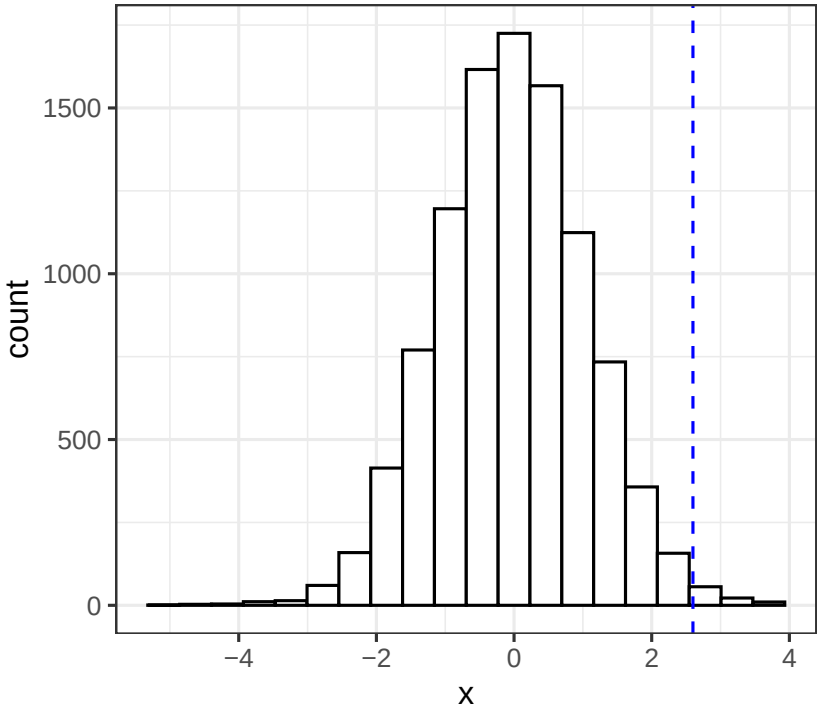
Ta áp dụng kiểm định bootstrap:

- số lần lặp bootstrap $B = 10,000$ lần;
- p-value tính bằng hai cách: hiệu số và pivot;

Trung bình hai mẫu - đồng phương sai



(a) hiệu số



(b) pivot

Xấp xỉ *p*-value

- bootstrap hiệu số: 0.0218
- bootstrap pivot: 0.0163

Trung bình hai mẫu - đồng phương sai

Trong trường hợp ta muốn kiểm tra giả thuyết một phía:

$$\begin{cases} H_0: & \mu_X \geq \mu_Y \\ H_1: & \mu_X < \mu_Y \end{cases} \qquad \text{hoặc} \qquad \begin{cases} H_0: & \mu_X \leq \mu_Y \\ H_1: & \mu_X > \mu_Y \end{cases}$$

Quy trình bootstrap sẽ là tương tự như trường hợp kiểm định hai phía.

p-value sẽ được tính bởi

- kiểm định một phía bên trái ($<$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(t_j^* < t_{\text{obs}}\right),$$

- kiểm định một phía bên phải ($>$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(t_j^* > t_{\text{obs}}\right),$$

với

$$t_{\text{obs}} = \overline{X} - \overline{Y}, \qquad \text{hoặc} \qquad t_{\text{obs}} = \frac{(\overline{X} - \overline{Y})}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}.$$

Trung bình hai mẫu - khác phương sai

Cho hai mẫu ngẫu nhiên

- $X_1, \ldots, X_{n_1} \overset{i.i.d}{\sim} F$;
- $Y_1, \ldots, Y_{n_2} \overset{i.i.d}{\sim} G$,

với $\mathbb{E}(X) = \mu_X$ và $\mathbb{E}(Y) = \mu_Y$.

Ta quan tâm tới kiểm định giả thuyết

$$H_0 : \mu_X = \mu_Y \qquad \text{vs.} \qquad H_1 : \mu_X \neq \mu_Y.$$

Và không muốn giả định gì thêm về phân phối hay phương sai.

Theo lý thuyết kiểm định, thống kê

$$T_{n_1,n_2} = \frac{(\overline{X} - \overline{Y})}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}} \xrightarrow{d} \mathcal{N}(0,1),$$

nếu H_0 đúng, với điều kiện $n_1 \to \infty, n_2 \to \infty$ và $n_1/n_2 \to \lambda$ khi $n_1 \to \infty$.

Trung bình hai mẫu - khác phương sai

p-value được xác định bởi

$$p\text{-value} = \Pr_{F_0}\left(\frac{|\overline{X}-\overline{Y}|}{\sqrt{\frac{S_X^2}{n_1}+\frac{S_Y^2}{n_2}}} > |t_{obs}|\right),$$

với phân phối F_0 được xác định là phân phối chuẩn nếu n_1, n_2 lớn.

Trong thực tế,

- tỷ lệ cỡ mẫu n_1/n_2 khó đảm bảo đặc biệt là cỡ mẫu lớn;
- sự xấp xỉ chuẩn còn phụ thuộc vào độ lệch của phân phối.

Do đó, việc áp dụng xấp xỉ chuẩn theo định lý giới hạn trung tâm không cho kết quả tốt.

Trung bình hai mẫu - khác phương sai

Ta áp dụng bootstrap để tạo mẫu phỏng theo giả thuyết H_0, tức là $\mu_X = \mu_Y$.

Vì

- X và Y không cùng phân phối;
- σ_X^2 và σ_Y^2 là không nhất thiết bằng nhau;

$\Rightarrow$ ta không thể ghép hai mẫu này thành một.

$\Rightarrow$ phải sử dụng hai mẫu: $X_1, \ldots, X_n$ và $Y_1, \ldots, Y_n$.

Gọi X^* và Y^* là hai mẫu bootstrap của X và Y.

Để đảm bảo mẫu bootstrap được xây dựng dưới giả thuyết H_0, ta cần đảm bảo rằng $\mathbb{E}(X^*) = \mathbb{E}(Y^*)$.

Nếu X^* và Y^* được tạo theo phân phối thực nghiệm $\widehat{F}_{n_1}$ và $\widehat{G}_{n_2}$ thì

$$\mathbb{E}_{\widehat{F}_{n_1}}(X^*) = \overline{X} \neq \overline{Y} = \mathbb{E}_{\widehat{G}_{n_2}}(Y^*).$$

Do đó, mẫu bootstrap X^* và Y^* sẽ được tạo bởi

$$X^* = X - \overline{X}, \qquad \text{và} \qquad Y^* = Y - \overline{Y}.$$

Trung bình hai mẫu - khác phương sai

Kiểm định bootstrap cho hai trung bình, khác phương sai

1. Ta tạo hai mẫu ngẫu nhiên cùng với cỡ của dữ liệu gốc từ dữ liệu gốc, có lặp lại (with replacement):
   $X_1^*, X_2^*, \ldots, X_{n_1}^*$ với $X_j^* = X_{j^*} - \overline{X}$;
   $Y_1^*, Y_2^*, \ldots, Y_{n_2}^*$ với $Y_j^* = Y_{j^*} - \overline{Y}$.
2. Tính ước lượng $\overline{X}^*$, S_X^{*2}, $\overline{Y}^*$ và S_Y^{*2} từ dữ liệu vừa tạo.
3. Tính giá trị thống kê

$$t^* = \overline{X}^* - \overline{Y}^*, \qquad \text{hoặc} \qquad t^* = \frac{(\overline{X}^* - \overline{Y}^*)}{\sqrt{\dfrac{S_X^{*2}}{n_1} + \dfrac{S_Y^{*2}}{n_2}}},$$

4. Lặp lại bước 1, 2 và 3 trong B lần (ít nhất 2000 lần), và lưu kết quả lại.

Trung bình hai mẫu - khác phương sai

5. Tính p-value theo công thức

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\left|t_j^*\right| \geq \left|t_{\text{obs}}\right|\right),$$

với

$$t_{\text{obs}} = \overline{X} - \overline{Y}, \qquad \text{hoặc} \qquad t_{\text{obs}} = \frac{(\overline{X} - \overline{Y})}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}}.$$

Trung bình hai mẫu - khác phương sai

Ví dụ: Một thí nghiệm được thực hiện bởi Catlin và Wang (2013) để kiểm tra xem liệu việc được tái chế giấy có khiến cho người sử dụng giấy không quan tâm tới lượng giấy họ sử dụng.

Hai nhóm thí nghiệm:

- nhóm 1: có thùng rác nhưng không phải loại rác có thể tái chế;
- nhóm 2: có thùng rác tái chế,

các hai nhóm được cung cấp giấy và kéo, và được yêu cầu ràng phải kiểm tra độ sắc của kéo.

Số tờ giấy được sử dụng được thu thập lại.

	n	mean	sd
Nhóm 1	22	9.4546	5.1337
Nhóm 2	19	17.2106	12.5547

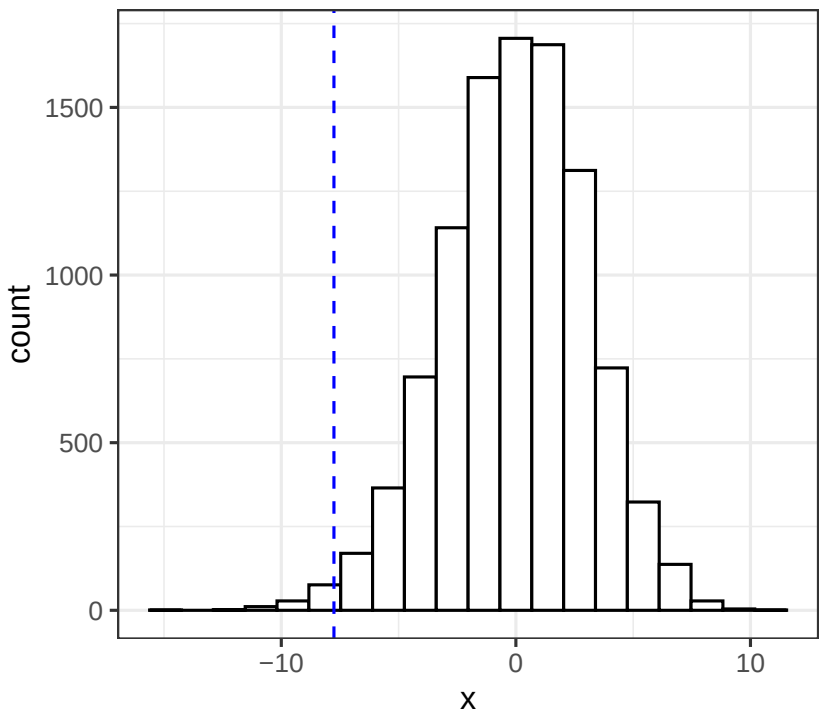
Nhận thấy rằng, tỷ số của hai độ lệch chuẩn là 0.41, không nằm trong (0.5, 2).

Do đó, ta không thể giả định rằng phương sai là đồng nhất giữa hai nhóm.

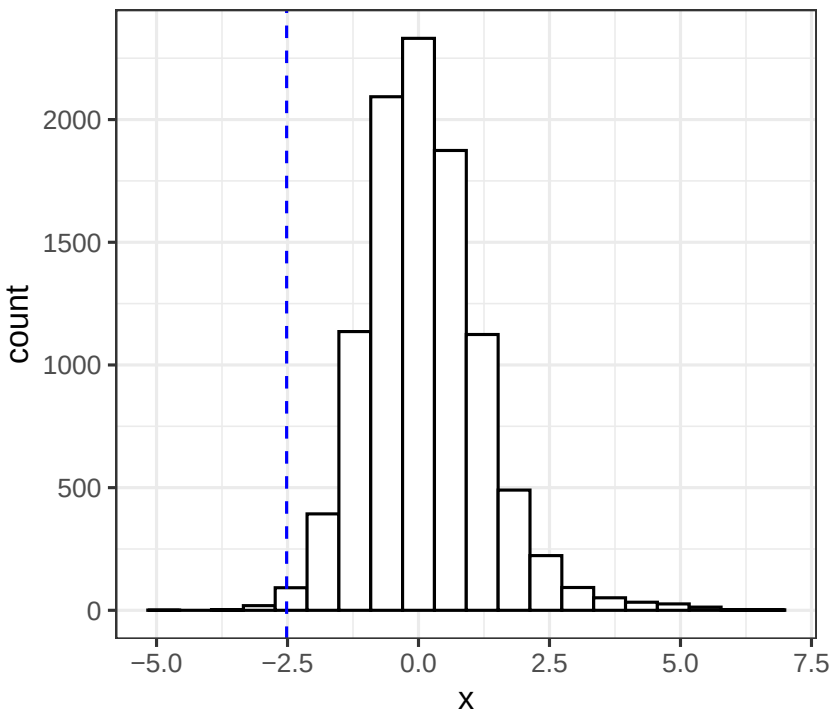
Ta áp dụng kiểm định bootstrap:

- số lần lặp bootstrap $B = 10,000$ lần;
- p-value tính bằng hai cách: hiệu số và pivot.

Trung bình hai mẫu - khác phương sai



(a) hiệu số



(b) pivot

Xấp xỉ *p*-value

- bootstrap hiệu số: 0.0118
- bootstrap pivot: 0.0327

Trung bình hai mẫu - khác phương sai

Trong trường hợp ta muốn kiểm tra giả thuyết một phía:

$$\begin{cases} H_0: & \mu_X \geq \mu_Y \\ H_1: & \mu_X < \mu_Y \end{cases} \qquad \text{hoặc} \qquad \begin{cases} H_0: & \mu_X \leq \mu_Y \\ H_1: & \mu_X > \mu_Y \end{cases}$$

Quy trình bootstrap sẽ là tương tự như trường hợp kiểm định hai phía.

p-value sẽ được tính bởi

- kiểm định một phía bên trái ($<$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(t_j^* < t_{\text{obs}}\right),$$

- kiểm định một phía bên phải ($>$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(t_j^* > t_{\text{obs}}\right),$$

với

$$t_{\text{obs}} = \overline{X} - \overline{Y}, \qquad \text{hoặc} \qquad t_{\text{obs}} = \frac{(\overline{X} - \overline{Y})}{\sqrt{\dfrac{S_X^2}{n_1} + \dfrac{S_Y^2}{n_2}}}.$$

Kiểm định sự tương quan tuyến tính

Cho bộ dữ liệu ngẫu nhiên $(X_1, Y_1), \ldots, (X_n, Y_n)$ của hai biến ngẫu nhiên X và Y.

Đặt r là hệ số tương quan Pearson giữa X và Y.

Ta quan tâm tới kiểm định sự độc lập (tuyến tính) giữa X và Y:

$$H_0 : r = 0 \qquad \text{vs.} \qquad H_1 : r \neq 0$$

Dưới giả định H_0, ta nói rằng X và Y là độc lập tuyến tính, tức là:

- X và Y có thể có tương quan phi tuyến;
- X và Y là độc lập.

Để kiểm định H_0, cách tiếp cận thông thường:

- giả sử (X, Y) tuân theo phân phối chuẩn 2 chiều;
- sử dụng thống kê kiểm định

$$\widehat{r}\sqrt{\frac{n-2}{1-\widehat{r}^2}} \sim t_{n-2}.$$

Kiểm định sự tương quan tuyến tính

Trong thực tế, việc giả định phân phối chuẩn 2 chiều là khó để kiểm chứng.

Mục tiêu là áp dụng tạo mẫu bootstrap dưới giả thuyết H_0. Nếu

- X và Y là có tương quan phi tuyến

  $\Rightarrow$ cần thông tin về sự tương quan tuyến tính;

  $\Rightarrow$ khó tiếp cận và không có tính tổng quát;
- X và Y là độc lập

  $\Rightarrow$ sử dụng tính chất $F_{X,Y}(x,y) = F_X(x)F_Y(y)$;

  $\Rightarrow$ tạo mẫu bootstrap độc lập theo ước lượng thực nghiệm của phân phối lề $\widehat{F}_{X,n}(x)$ và $\widehat{F}_{Y,n}(y)$.

Kiểm định sự tương quan tuyến tính

Giả sử ta có mẫu

- $X_1^*, X_2^*, \ldots, X_n^*$ tạo từ $\widehat{F}_{X,n}$;
- $Y_1^*, Y_2^*, \ldots, Y_n^*$ tạo từ $\widehat{F}_{Y,n}$.

Do hai mẫu này được tạo độc lập nhau $\Rightarrow$ ước lượng hệ số tương quan Pearson $\widehat{r}^*$ sẽ có giá trị xung quanh 0.

Giả sử ta có B mẫu bootstrap được tạo như trên:

- thu được phân phối mẫu của $\widehat{r}^*$ dưới giả định H_0, tức là $\mathbb{E}(\widehat{r}^*) = 0$;
- nếu H_0 thực sự đúng thì $\widehat{r}$ sẽ nằm trong vùng giữa của phân phối mẫu bootstrap;
- nếu mẫu không tương thích với H_0 thì $\widehat{r}$ sẽ nằm ở một trong hai phần đuôi của phân phối mẫu bootstrap.

Từ đây, ta định nghĩa p-value bởi

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\left|\widehat{r}_j^*\right| \geq |\widehat{r}|\right).$$

Kiểm định sự tương quan tuyến tính

Kiểm định bootstrap cho tương quan tuyến tính

1. Ta tạo hai mẫu ngẫu nhiên cùng với cỡ của dữ liệu gốc từ dữ liệu gốc, có lặp lại (with replacement):
   $X_1^*, X_2^*, \ldots, X_{n_1}^*$;
   $Y_1^*, Y_2^*, \ldots, Y_{n_2}^*$.
2. Tính ước lượng $\widehat{r}^*$ và $V(\widehat{r}_j^*)$ từ dữ liệu vừa tạo.
3. Lặp lại bước 1 và 2 trong B lần (ít nhất 2000 lần), và lưu kết quả lại.
4. Tính p-value theo công thức

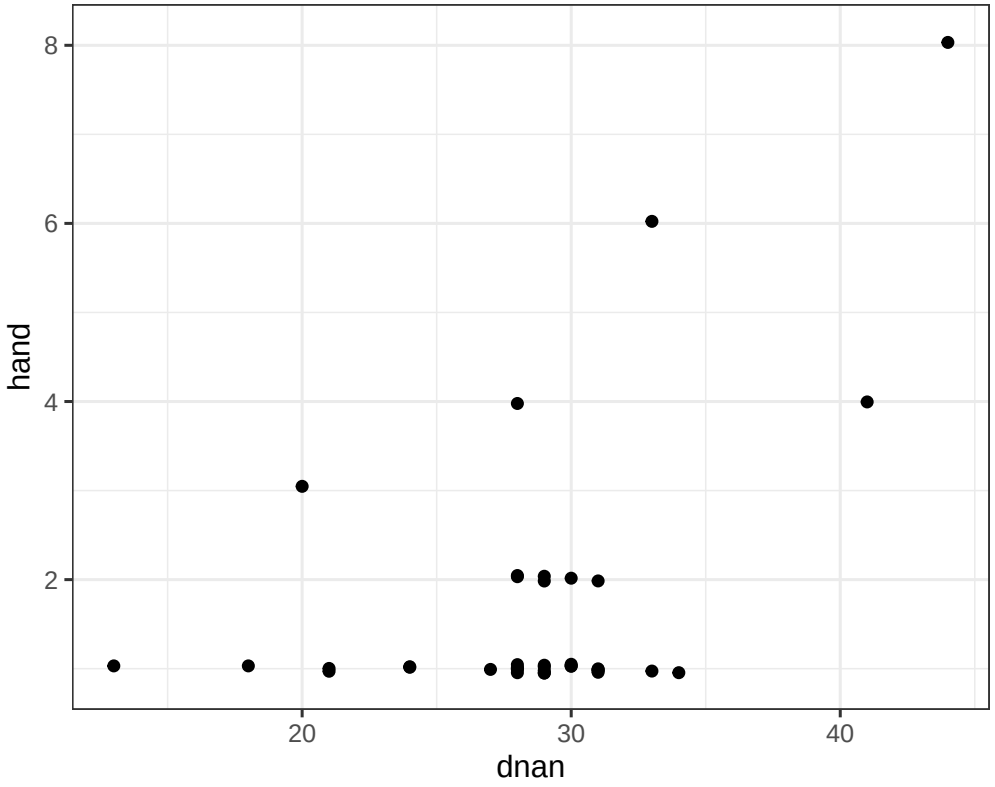
$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\left|\widehat{r}_j^*\right| \geq |\widehat{r}|\right),$$

hoặc sử dụng cách tiếp cận pivot

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\left|\frac{\widehat{r}_j^*}{\sqrt{V(\widehat{r}_j^*)}}\right| \geq \left|\frac{\widehat{r}}{\sqrt{V(\widehat{r})}}\right|\right).$$

Kiểm định sự tương quan tuyến tính

Ví dụ: ta xét dữ liệu về nghiên cứu về mối liên hệ giữa khiếm khuyết gen và tay thuận của một người:



Hệ số tương quan tuyến tính (Pearson), $\widehat{r} = 0.509$.

Kiểm định sự tương quan tuyến tính

Ta muốn kiểm tra xem liệu dữ liệu có thực sự tương quan tuyến tính.

$$H_0 : r = 0 \qquad \text{vs.} \qquad H_1 : r \neq 0$$

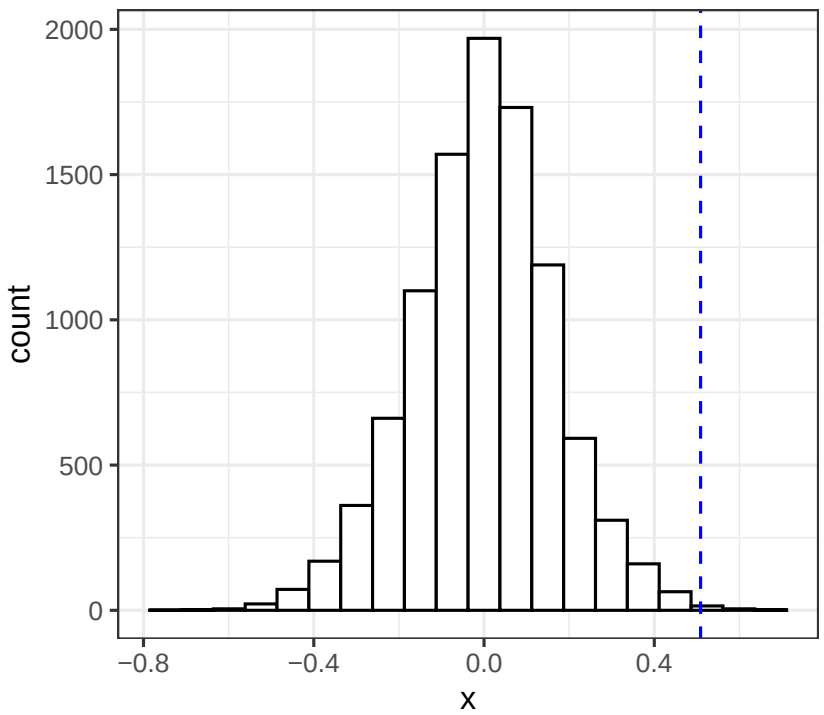
Rõ ràng, việc giả định dữ liệu phân phối chuẩn hai chiều là không hợp lý.

$\Rightarrow$ việc sử dụng cách kiểm định thông qua phân phối t-student cho kết quả không đáng tin cậy.

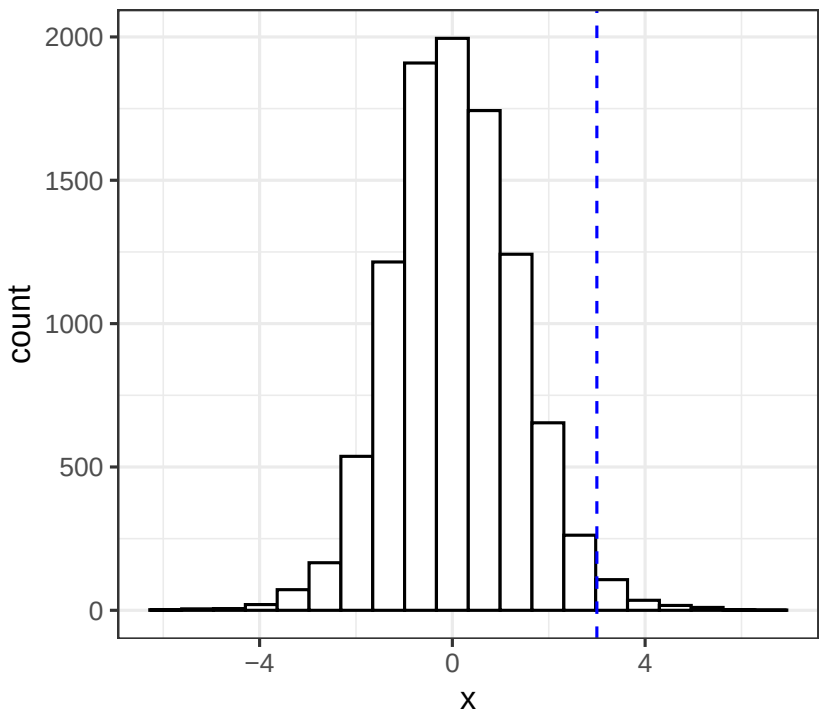
Ta áp dụng kiểm định bootstrap:

- số lần lặp bootstrap $B = 10,000$ lần;
- p-value tính bằng hai cách: so sánh và pivot;
- phương sai của ước lượng được ước lượng bằng phương pháp influence function;
- $\widehat{r} = 0.509,\ V(\widehat{r}) = 0.029 \Rightarrow \dfrac{\widehat{r}}{\sqrt{V(\widehat{r})}} = 2.9985$

Kiểm định sự tương quan tuyến tính



(a) $\widehat{r}^*$



(b) pivot

Xấp xỉ p-value

- bootstrap: 0.004
- bootstrap pivot: 0.027

Kiểm định sự tương quan tuyến tính

Trong trường hợp ta muốn kiểm tra giả thuyết một phía:

$$\begin{cases} H_0: & r \geq 0 \\ H_1: & r < 0 \end{cases} \quad \text{hoặc} \quad \begin{cases} H_0: & r \leq 0 \\ H_1: & r > 0 \end{cases}$$

Quy trình bootstrap sẽ là tương tự như trường hợp kiểm định hai phía.

p-value sẽ được tính bởi

- kiểm định một phía bên trái ($<$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\widehat{r}_j^* < \widehat{r}\right), \text{ hoặc } p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\frac{\widehat{r}_j^*}{\sqrt{V(\widehat{r}_j^*)}} < \frac{\widehat{r}}{\sqrt{V(\widehat{r})}}\right).$$

- kiểm định một phía bên phải ($>$)

$$p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\widehat{r}_j^* > \widehat{r}\right) \text{ hoặc } p\text{-value} = \frac{1}{B}\sum_{j=1}^{B} \mathrm{I}\left(\frac{\widehat{r}_j^*}{\sqrt{V(\widehat{r}_j^*)}} > \frac{\widehat{r}}{\sqrt{V(\widehat{r})}}\right).$$

Kiểm định cho tỷ số trung bình

Cho hai dữ liệu ngẫu nhiên $X_1, \ldots, X_{n_1}$ và $Y_1, \ldots, Y_{n_2}$ của hai biến ngẫu nhiên X và Y.

Đặt θ là tỷ số trung bình giữa X và Y.

Ta quan tâm tới kiểm định:

$$H_0 : \theta = 1 \qquad \text{vs.} \qquad H_1 : \theta \neq 1$$

Dưới giả định H_0, ta nói rằng X và Y có trung bình bằng nhau.

Hãy xây dựng một quy trình kiểm định bootstrap cho kiểm định này.