THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 26 tháng 9 năm 2018

Nội dung môn học

1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. Kiểm định giả thuyết về tham số một tổng thể
8. **Kiểm định giả thuyết về tham số hai tổng thể**
9. Phân tích phương sai
10. Kiểm định phi tham số
11. Kiểm định chi - bình phương

Phần VII

Kiểm định giả thiết tham số hai tổng thể

Mục lục

Example

Một công ty điều tra thị trường lao động công bố rằng thu nhập trung bình của những sinh viên sau 10 năm tốt nghiệp TLU cao hơn so với thu nhập của sinh viên tương ứng tốt nghiệp HUST. Ở một khía cạch khác, họ cho rằng sinh viên TLU khi ra trường có mức thu nhập đồng đều hơn, tuy vậy, tỉ lệ xin được việc đúng ngành của sinh viên TLU thấp hơn so với HUST. Làm sao để bạn kiểm định lại những điều này?

Solution

1. *Bài toán thứ nhất* $\longrightarrow$ *So sánh hai trung bình.*
2. *Bài toán thứ hai* $\longrightarrow$ *So sánh hai phương sai.*
3. *Bài toán thứ ba* $\longrightarrow$ *So sánh hai tỉ lệ.*

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Bài toán

Hai tổng thể với trung bình lần lượt là μ_1, μ_2. Ta cần so sánh hai trung bình μ_1 và μ_2 chênh lệch so với nhau một lượng D_0 dựa trên việc kiểm định các cặp giả thuyết H_0, H_1 sau:

	Bài toán bên phải	*Bài toán bên trái*	*Bài toán hai bên*
$H_0:$	$\mu_1 - \mu_2 = D_0$	$\mu_1 - \mu_2 = D_0$	$\mu_1 - \mu_2 = D_0$
	$\mu_1 - \mu_2 \leq D_0$	$\mu_1 - \mu_2 \geq D_0$	
$H_1:$	$\mu_1 - \mu_2 > D_0$	$\mu_1 - \mu_2 < D_0$	$\mu_1 - \mu_2 \neq D_0$

Solution

Thống kê sử dụng để giải quyết các bài toán trên phụ thuộc vào ý tưởng chọn mẫu: độc lập hay theo cặp.

Chọn mẫu độc lập

Định nghĩa

Khi so sánh trung bình hai tổng thể, tùy theo cách chọn mẫu mà sử dụng quy luật phân phối khác nhau:

1. *Chọn hai mẫu độc lập.*
2. *Chọn hai mẫu theo đôi*

Chọn mẫu độc lập

Định nghĩa

Khi so sánh trung bình hai tổng thể, tùy theo cách chọn mẫu mà sử dụng quy luật phân phối khác nhau:

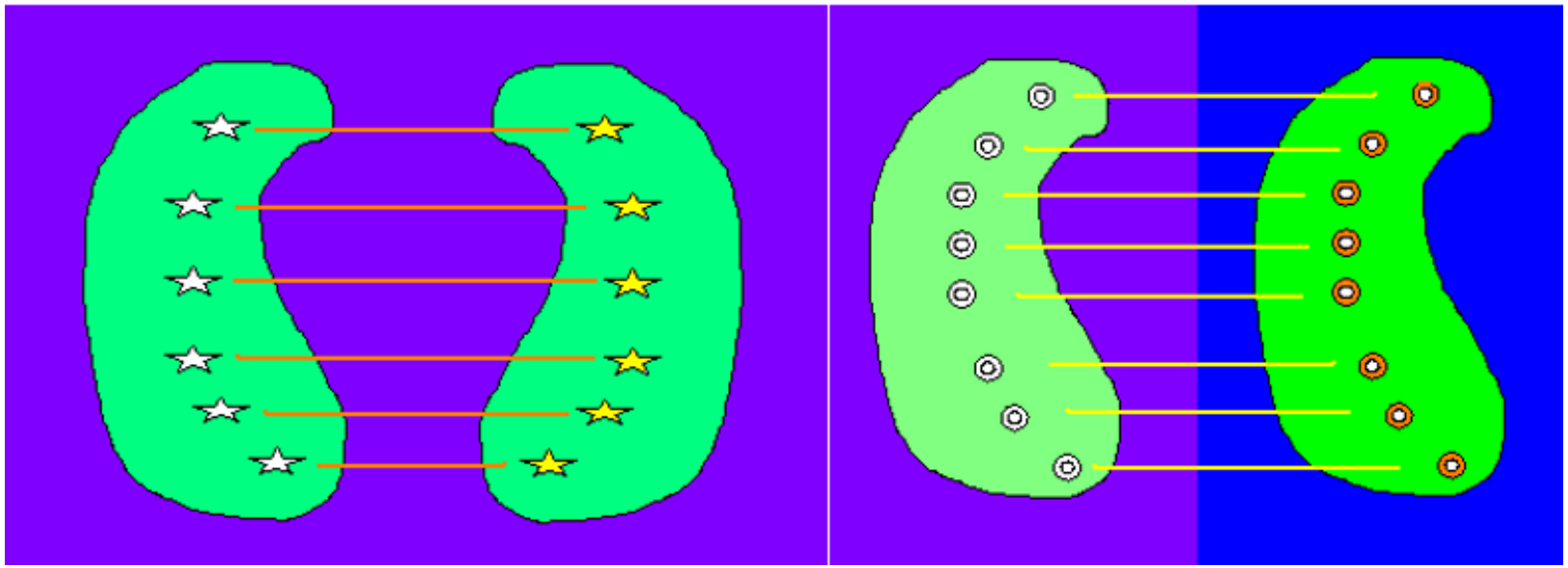
1. *Chọn hai mẫu độc lập.*
2. *Chọn hai mẫu theo đôi*

Hai mẫu được chọn **độc lập** *là các mẫu được chọn từ mỗi tổng thể sao cho một quan sát được chọn vào mẫu này không ảnh hưởng đến xác suất một quan sát nào đó được chọn vào mẫu kia.*

Example

Chẳng hạn, để so sánh mức độ luộm thuộm của những người nghiện thuốc lá và nghiện rượu, ta chọn ngẫu nhiên 20 người nghiện thuốc lá và chọn **ngẫu nhiên** 30 người nghiện rượu. Hai mẫu này độc lập. Còn nếu ta chọn 30 người nghiện rượu phải là bạn bè của 20 người nghiện thuốc lá thì không phải là chọn hai mẫu độc lập.

Chọn mẫu theo đôi



- mỗi đối tượng của hai mẫu được chọn sao cho tương ứng với nhau căn cứ vào tiêu chuẩn nào đó, hoặc:
- đối tượng ở mẫu thứ hai cũng chính là đối tượng ở mẫu thứ nhất

Example

Để nghiên cứu tác dụng của các đổi mới quy trình tiếp dân đối với thái độ phục vụ của cán bộ phường:
Cách thức chọn mẫu: khảo sát thái độ phục vụ tại 20 phường trước khi ban hành quy trình mới. Sau khi ban hành các quy trình mới, ta lại khảo sát thái độ phục vụ của cán bộ tại chính 20 phường trên.

Một số ký hiệu

Quy luật bác bỏ và chấp nhận H_0

- Gọi μ_1, μ_2 lần lượt là trung bình của tổng thể thứ nhất và thứ hai;
- Gọi σ_1^2, σ_2^2 lần lượt là phương sai của tổng thể thứ nhất và thứ hai;
- Gọi s_1^2, s_2^2 lần lượt là phương sai của mẫu thứ nhất và thứ hai;
- Gọi $\overline{X}_1, \overline{X}_2$ lần lượt là biến ngẫu nhiên chỉ trung bình của mẫu được chọn từ tổng thể thứ nhất và tổng thể thứ 2;
- Gọi $\bar{x}_1, \bar{x}_2$ lần lượt là giá trị của $\overline{X}_1, \overline{X}_2$ tính trên mẫu cụ thể mà ta chọn.

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Theorem

Giả sử:

1. *hai tổng thể cần nghiên cứu tuân theo phân phối chuẩn với trung bình lần lượt là* μ_1, μ_2 *chưa biết.*
2. *phương sai của hai tổng thể* σ_1^2, σ_2^2 *đã biết.*
3. *hai mẫu chọn ra từ hai tổng thể theo cách độc lập nhau.*

Khi đó ta có kết quả:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Example

Một trong những chỉ tiêu so sánh chất lượng phục vụ bay là thời gian delay bay trung bình của hãng đó. Giả sử ta có dữ liệu thời gian (giờ) delay của một số chuyến bay của hãng A (20 chuyến), B(25 chuyến) như sau:

A 3.5 2.8 2.3 2.8 1.4 1.7 2.8 3.9 3.6 4.0 0.4 1.5 1.1 2.3 4.3 2.9 2.0 4.3 2.1 2.6

B 2.7 4.0 5.5 2.6 5.6 0.9 4.3 3.8 4.0 4.6 3.5 4.8 2.5 2.6 3.7 5.3 4.0 2.9 2.9 3.8 5.6 2.7 3.1 5.7 5.0

Giả sử biết độ lệch chuẩn của thời gian hoãn bay của hai hãng A, B lần lượt là 1 và 1.5 (giờ) và cả hai tổng thể đều có phân phối chuẩn.
Hỏi, tại mức ý nghĩa 5% thời gian hoãn bay trung bình của hai hãng có như nhau không?

Solution

Gọi μ_1, μ_2 lần lượt là thời gian hoãn bay trung bình của hãng A và B.

- $H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0.$
- *Giả sử H_0 đúng:*

Solution

Gọi μ_1, μ_2 lần lượt là thời gian hoãn bay trung bình của hãng A và B.

- $H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0.$
- *Giả sử H_0 đúng: Khi đó đại lượng* $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\dfrac{1}{20} + \dfrac{1.5^2}{25}}} \sim N(0, 1).$

Solution

Gọi μ_1, μ_2 lần lượt là thời gian hoãn bay trung bình của hãng A và B.

- $H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0.$
- *Giả sử H_0 đúng: Khi đó đại lượng* $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\dfrac{1}{20} + \dfrac{1.5^2}{25}}} \sim N(0,1).$

  Do đó, nếu H_0 đúng thì giá trị Z tính được từ mẫu sẽ nằm ngoài khoảng $[-1.96; 1.95]$ chỉ chiếm 5%.

Solution

Gọi μ_1, μ_2 lần lượt là thời gian hoãn bay trung bình của hãng A và B.

- $H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0.$
- *Giả sử H_0 đúng: Khi đó đại lượng* $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\dfrac{1}{20} + \dfrac{1.5^2}{25}}} \sim N(0,1).$

  Do đó, nếu H_0 đúng thì giá trị Z tính được từ mẫu sẽ nằm ngoài khoảng $[-1.96; 1.95]$ chỉ chiếm 5%.
  Ta tính được $\bar{x_1} = 2.615, \bar{x_2} = 3.844$ nên $z = -3.2846$ nằm ngoài khoảng nói trên. Do vậy, với sai lầm không quá 5% ta có thể bác bỏ H_0.

Solution

Gọi μ_1, μ_2 lần lượt là thời gian hoãn bay trung bình của hãng A và B.

- $H_0 : \mu_1 - \mu_2 = 0 \qquad H_1 : \mu_1 - \mu_2 \neq 0.$
- *Giả sử H_0 đúng: Khi đó đại lượng* $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{\dfrac{1}{20} + \dfrac{1.5^2}{25}}} \sim N(0,1).$

  Do đó, nếu H_0 đúng thì giá trị Z tính được từ mẫu sẽ nằm ngoài khoảng $[-1.96; 1.95]$ chỉ chiếm 5%.
  Ta tính được $\bar{x_1} = 2.615, \bar{x_2} = 3.844$ nên $z = -3.2846$ nằm ngoài khoảng nói trên. Do vậy, với sai lầm không quá 5% ta có thể bác bỏ H_0.
- *Vậy, thời gian hoãn bay trung bình của hai hãng là khác nhau.*

Quy luật bác bỏ và chấp nhận H_0

H_0	H_1	Thống kê z	Qui luật bác bỏ H_0	p-giá trị
$\mu_1 - \mu_2 \leq D_0$	$\mu_1 - \mu_2 > D_0$	$\dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$z > z_\alpha$	$P(Z > z)$
$\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 < D_0$		$z < -z_\alpha$	$P(Z < z)$
$\mu_1 - \mu_2 = D_0$	$\mu_1 - \mu_2 \neq D_0$		$\lvert z \rvert > z_{\alpha/2}$	$2P(Z > z)$

Example

Nếu dùng P - giá trị, ở ví dụ trên ta có thể tính:

$$\text{P - giá trị} = 2P(Z > |z|) = 2P(Z > 3.248) =$$
$$2 * pnorm(3.248, lower.tail = FALSE) = 0.001162193$$

Ta thấy, P - giá trị < 0.05 nên ta có thể bác bỏ H_0 với sai lầm không quá 5%.

Trên R

Trên R, để so sánh trung bình hai tổng thể có phân phối chuẩn, biết phương sai tương ứng σ_1^2, σ_2^2, mẫu chọn độc lập ta dùng:

$$z.test(\text{Mẫu 1}, \text{Mẫu 2}, sigma.x = \sigma_1, sigma.y = \sigma_2, mu = D_0, alt)$$

trong đó, alt = "t" nếu là bài toán hai bên, "less" nếu là bài toán bên trái, "g" nếu là bài toán bên phải.
Lưu ý, hàm z.test không có sẵn trong R bản cơ sở nên nếu cần dùng phải cài thêm gói, chẳng hạn BSDA bằng cách gõ

$$install.packages(\text{"}BSDA\text{"})$$

và chọn đường link tải hiện ra (nếu có).
Trong mỗi lần dùng ta phải gõ

$$library(BSDA)$$

Example

Chẳng hạn đối với bài trên, ta dùng dãy lệnh:

```
> library(BSDA)
> A=scan()
1: 3.5 2.8 2.3 2.8 1.4 1.7 2.8 3.9 3.6 4.0 0.4 1.5 1.1 2.3 4.3 2.9 2
.0 4.3 2.1 2.6
21:
Read 20 items
> B=scan()
1: 2.7 4.0 5.5 2.6 5.6 0.9 4.3 3.8 4.0 4.6 3.5 4.8 2.5 2.6 3.7 5.3 4
.0 2.9 2.9 3.8 5.6 2.7 3.1 5.7 5.0
26:
Read 25 items
> z.test(A,B,alt="t",mu=0,sigma.x = 1,sigma.y = 1.5)
```

Kết quả cho ta $p - value = 0.001021 < 0.05$ nên bác bỏ H_0, từ đó kết luận về bài toán.

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Theorem

Nếu mẫu được chọn ra từ hai tổng thể với cỡ lớn, $n_1 \geq 30, n_2 \geq 30$, phương sai σ_1^2, σ_2^2 của hai tổng thể chưa biết có thể thay lần lượt bởi các phương sai mẫu S_1^2, S_2^2. Khi đó biến ngẫu nhiên

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

xấp xỉ phân phối chuẩn hóa.

Example

Nhiều ý kiến cho rằng lương của phụ nữ thấp hơn lương nam giới. Để kiểm định điều này, người ta tiến hành điều tra 100 nam giới thì thấy lương trung bình là 7 (triệu/tháng) với độ lệch chuẩn là 2, điều tra 90 phụ nữ thấy lương trung bình là 6.3, độ lệch chuẩn là 1.5. Ở mức ý nghĩa $\alpha = 5\%$ hãy hãy kiểm định ý kiến trên.

Solution

Gọi μ_1, μ_2 lần lượt là lương trung bình của nam giới và nữ giới.

1. *Cặp giả thuyết:*

Example

Nhiều ý kiến cho rằng lương của phụ nữ thấp hơn lương nam giới. Để kiểm định điều này, người ta tiến hành điều tra 100 nam giới thì thấy lương trung bình là 7 (triệu/tháng) với độ lệch chuẩn là 2, điều tra 90 phụ nữ thấy lương trung bình là 6.3, độ lệch chuẩn là 1.5. Ở mức ý nghĩa $\alpha = 5\%$ hãy hãy kiểm định ý kiến trên.

Solution

Gọi μ_1, μ_2 lần lượt là lương trung bình của nam giới và nữ giới.

1. *Cặp giả thuyết:* $H_0 : \mu_1 - \mu_2 \leq 0 \qquad H_1 : \mu_1 - \mu_2 > 0.$

Example

Nhiều ý kiến cho rằng lương của phụ nữ thấp hơn lương nam giới. Để kiểm định điều này, người ta tiến hành điều tra 100 nam giới thì thấy lương trung bình là 7 (triệu/tháng) với độ lệch chuẩn là 2, điều tra 90 phụ nữ thấy lương trung bình là 6.3, độ lệch chuẩn là 1.5. Ở mức ý nghĩa $\alpha = 5\%$ hãy hãy kiểm định ý kiến trên.

Solution

Gọi μ_1, μ_2 lần lượt là lương trung bình của nam giới và nữ giới.

1. *Cặp giả thuyết:* $H_0 : \mu_1 - \mu_2 \leq 0 \qquad H_1 : \mu_1 - \mu_2 > 0.$
2. *Ta có* $z = \dfrac{(7 - 6.3) - 0}{\sqrt{\dfrac{2^2}{100} + \dfrac{1.5^2}{90}}} = 2.7456.$

Example

Nhiều ý kiến cho rằng lương của phụ nữ thấp hơn lương nam giới. Để kiểm định điều này, người ta tiến hành điều tra 100 nam giới thì thấy lương trung bình là 7 (triệu/tháng) với độ lệch chuẩn là 2, điều tra 90 phụ nữ thấy lương trung bình là 6.3, độ lệch chuẩn là 1.5. Ở mức ý nghĩa $\alpha = 5\%$ hãy hãy kiểm định ý kiến trên.

Solution

Gọi μ_1, μ_2 lần lượt là lương trung bình của nam giới và nữ giới.

1. *Cặp giả thuyết:* $H_0 : \mu_1 - \mu_2 \leq 0 \qquad H_1 : \mu_1 - \mu_2 > 0.$
2. *Ta có* $z = \dfrac{(7 - 6.3) - 0}{\sqrt{\dfrac{2^2}{100} + \dfrac{1.5^2}{90}}} = 2.7456.$

   P - giá trị$= P(Z > 2.7456) = 1 - P(Z \leq 2.7456) = 0.00302.$
3. *Ta có P - giá trị* < 0.05 *nên bác bỏ* H_0.
4. *Vậy, ở mức ý nghĩa 5%, lương trung bình của nam giới cao hơn lương trung bình của nữ giới.*

Trên R

Khi cỡ mẫu lớn, ta xấp xỉ

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$

So sánh với trường hợp biết σ_1, σ_2:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Ta thấy, σ_1, σ_2 được thay thế bởi S_1, S_2 tương ứng nên ta vẫn có thể dùng

$$z.test(\text{Mẫu 1}, \text{Mẫu 2}, sigma.x = s_1, sigma.y = s_2, mu = D_0, alt)$$

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - **Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập**
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Theorem

Giả sử hai tổng thể nghiên cứu tuân theo phân phối chuẩn với trung bình lần lượt là μ_1, μ_2 chưa biết và phương sai σ_1^2, σ_2^2 chưa biết nhưng không có giả định bằng nhau: $\sigma_1^2 \neq \sigma_2^2$. Khi mẫu được chọn ra từ mỗi tổng thể này độc lập với nhau. Ta có

$$t_\upsilon = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

tuân theo phân phối student với bậc tự do υ được cho bởi công thức

$$\upsilon = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_2 - 1}}.$$

Quy luật bác bỏ và chấp nhận H_0

H_0	H_1	Thống kê t_υ	Qui luật bác bỏ H_0	p-giá trị
$\mu_1 - \mu_2 \leq D_0$	$\mu_1 - \mu_2 > D_0$	$\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	$t > t_{\upsilon,\alpha}$	$P(t_\upsilon > t)$
$\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 < D_0$		$t < -t_{\upsilon,\alpha}$	$P(t_\upsilon < t)$
$\mu_1 - \mu_2 = D_0$	$\mu_1 - \mu_2 \neq D_0$		$\lvert t\rvert > t_{\upsilon,\alpha/2}$	$2P(t_\upsilon > \lvert t\rvert)$

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - **Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập**
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Theorem

Giả sử hai tổng thể cần nghiên cứu tuân theo phân phối chuẩn. Khi phương sai của hai tổng thể chưa biết nhưng giả định $\sigma_1^2 = \sigma_2^2 = \sigma^2$, *khi đó đại lượng*

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

tuân theo phân phối Student với $n_1 + n_2 - 2$ *bậc tự do.*
Để cho gọn, ta kí hiệu

$$S = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}.$$

và

$$n = n_1 + n_2$$

Quy luật bác bỏ và chấp nhận H_0

H_0	H_1	Thống kê t	Qui luật bác bỏ H_0	p-giá trị
$\mu_1 - \mu_2 \leq D_0$	$\mu_1 - \mu_2 > D_0$	$\dfrac{(\overline{x}_1 - \overline{x}_2) - D_0}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$t > t_{n-2,\alpha}$	$P(t_{n-2} > t)$
$\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 < D_0$		$t < -t_{n-2,\alpha}$	$P(t_{n-2} < t)$
$\mu_1 - \mu_2 = D_0$	$\mu_1 - \mu_2 \neq D_0$		$\lvert t\rvert > t_{n-2,\alpha/2}$	$2P(t_{n-2} > \lvert t\rvert)$

Example

Một chủ chuỗi cửa hàng thời trang thử nghiệm để so sánh hiệu quả hai hình thức khuyến mãi tại 20 cửa hàng của mình. Nhóm 10 cửa hàng thứ nhất chạy khuyến mãi theo hình thức mua 1 tặng 1. Nhóm thứ hai theo hình thức giảm giá 50%. Sau một tuần, lợi nhuận (triệu đồng) tại 20 cửa hàng trên như sau:

- **Nhóm thứ nhất:**

$$7\ 10\ 9\ 8\ 6\ 12\ 10\ 7\ 10\ 7$$

- **Nhóm thứ hai:**

$$9\ 13\ 11\ 7\ 10\ 12\ 8\ 10\ 11\ 8$$

Giả sử rằng hai tổng thể tuân theo phân phối chuẩn với phương sai như nhau. Kiểm định sự khác biệt về hiệu quả của hai hình thức khuyến mãi trên. Chọn mức ý nghĩa 5%.

Solution

Gọi μ_1, μ_2 lần lượt là trung bình lợi nhuận của tổng thể các cửa hàng chạy chương trình khuyến mãi theo hình thức mua một tặng một, và của chương trình giảm giá 50%.

1. *Cặp giả thiết:*

Solution

Gọi μ_1, μ_2 lần lượt là trung bình lợi nhuận của tổng thể các cửa hàng chạy chương trình khuyến mãi theo hình thức mua một tặng một, và của chương trình giảm giá 50%.

1. *Cặp giả thiết:* $H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$.
2. *Ta có* $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{S\sqrt{1/10 + 1/10}} \sim t(18)$.

Solution

Gọi μ_1, μ_2 lần lượt là trung bình lợi nhuận của tổng thể các cửa hàng chạy chương trình khuyến mãi theo hình thức mua một tặng một, và của chương trình giảm giá 50%.

1. *Cặp giả thiết:* $H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$.
2. *Ta có* $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{S\sqrt{1/10 + 1/10}} \sim t(18)$. *Ta tính được*
   $s_1 = 1.897367, s_2 = 1.911951 \Rightarrow t = -1.5262.$

Solution

Gọi μ_1, μ_2 lần lượt là trung bình lợi nhuận của tổng thể các cửa hàng chạy chương trình khuyến mãi theo hình thức mua một tặng một, và của chương trình giảm giá 50%.

1. *Cặp giả thiết:* $H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$.
2. *Ta có* $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{S\sqrt{1/10 + 1/10}} \sim t(18)$. *Ta tính được* $s_1 = 1.897367, s_2 = 1.911951 \Rightarrow t = -1.5262$.
3. *Vùng chấp nhận là* $[-t_{18,0.025} = -2.1; t_{18,0.025} = 2.1]$. *Bởi vậy ta chấp nhận* H_0.

Solution

Gọi μ_1, μ_2 lần lượt là trung bình lợi nhuận của tổng thể các cửa hàng chạy chương trình khuyến mãi theo hình thức mua một tặng một, và của chương trình giảm giá 50%.

1. *Cặp giả thiết:* $H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$.
2. *Ta có* $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - 0}{S\sqrt{1/10 + 1/10}} \sim t(18)$. *Ta tính được* $s_1 = 1.897367, s_2 = 1.911951 \Rightarrow t = -1.5262$.
3. *Vùng chấp nhận là* $[-t_{18,0.025} = -2.1; t_{18,0.025} = 2.1]$. *Bởi vậy ta chấp nhận* H_0.
   (Nếu tính P - giá trị $= 2P(t_{18} > |1.5262|) = 0.1443 > 0.05$ *nên chấp nhận* H_0*)*
4. *Vậy, tại mức ý nghĩa 5%, không đủ bằng chứng thống kê đề cho rằng hai hình thức khuyến mãi đem đến hiệu quả khác nhau.*

Trong R

Đối với các kiểm đinh dùng phân phối t, ta dùng hàm

$$t.test(\text{Mẫu 1}, \text{Mẫu 2}, mu = D_0, alt, var.equal = TRUE/FALSE)$$

Trong đó, var.equal = TRUE nếu có giả thuyết hai phương sai là như nhau, = FALSE nếu chưa có giả thiết phương sai bằng nhau.

Example

Một chủ chuỗi cửa hàng thời trang thử nghiệm để so sánh hiệu quả hai hình thức khuyến mãi tại 20 cửa hàng của mình. Nhóm 10 cửa hàng thứ nhất chạy khuyến mãi theo hình thức mua 1 tặng 1. Nhóm thứ hai theo hình thức giảm giá 50%. Sau một tuần, lợi nhuận (triệu đồng) tại 20 cửa hàng trên như sau:

- **Nhóm thứ nhất:**

$$7\ 10\ 9\ 8\ 6\ 12\ 10\ 7\ 10\ 7$$

- **Nhóm thứ hai:**

$$9\ 13\ 11\ 7\ 10\ 12\ 8\ 10\ 11\ 8$$

Giả sử rằng hai tổng thể tuân theo phân phối chuẩn. Kiểm định sự khác biệt về hiệu quả của hai hình thức khuyến mãi trên trong hai trường hợp:

1. Phương sai hai tổng thể là như nhau.
2. Chưa có thông tin gì về phương sai hai tổng thể.

Chọn mức ý nghĩa 5%.

Solution (Trường hợp 1: Phương sai hai tổng thể là như nhau)

```
> T=scan()
1: 7 10  9  8  6 12 10  7 10  7
11:
Read 10 items
> G=scan()
1: 9 13 11  7 10 12  8 10 11  8
11:
Read 10 items
> t.test(T,G,alt="t",mu=0,var.equal = TRUE)

        Two Sample t-test

data:  T and G
t = -1.5262, df = 18, p-value = 0.1443
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0895559  0.4895559
sample estimates:
mean of x mean of y
      8.6       9.9
```

Ta được P - giá trị $= 0.1443 > 0.05$ nên chấp nhận H_0.

Solution (Trường hợp 1: Phương sai hai tổng thể là khác nhau)

```
> t.test(T,G,alt="t",mu=0,var.equal = FALSE)

        Welch Two Sample t-test

data:  T and G
t = -1.5262, df = 17.999, p-value = 0.1443
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.0895634  0.4895634
sample estimates:
mean of x mean of y
      8.6       9.9
```

Ta được P - giá trị $= 0.1443 > 0.05$ nên chấp nhận H_0.

Tổng kết các trường hợp khi mẫu chọn ra độc lập

Trường hợp	Thống kê sử dụng
Hai tổng thể PPC, biết phương sai	$\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$
Hai tổng thể PP bất kì, cỡ hai mẫu lớn	$\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \approx N(0,1)$
Hai tổng thể PPC, chưa biết PS $\sigma_1^2 \neq \sigma_2^2$	$\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \sim t(\upsilon)$
Hai tổng thể PPC, chưa biết PS $\sigma_1^2 = \sigma_2^2$	$\dfrac{(\overline{X}_1 - \overline{X}_2) - D_0}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n-2)$

Nhận xét

1. Thực tế ta không biết σ_1, σ_2. Nên trường hợp đầu ít dùng.
2. Khi tổng thể có phân phối chuẩn, ta có thể kiểm định cỡ mẫu nhỏ bằng kiểm định t, giả định về phương sai dẫn đến khác biệt ở bậc tự do.
3. Khi cỡ mẫu lớn ta dùng xấp xỉ phân phối N(0,1). Nhưng khi đó, ta cũng có thể dùng kiểm định t trong trường hợp phương sai khác nhau (vì khi đó công thức giá trị kiểm định giống nhau, phân phối $t(\upsilon) \approx N(0,1)$ do υ lớn)

Bởi thế, thực tế chủ yếu người ta dùng kiểm định t.

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Mẫu theo đôi

Chọn mẫu ngẫu nhiên gồm n cặp (X_{11}, X_{21}), $(X_{12}, X_{22}), \ldots,$ (X_{1n}, X_{2n}), trong đó $X_{11}, X_{12}, \ldots, X_{1n}$ là mẫu của tổng thể thứ nhất; $X_{21}, X_{22}, \ldots, X_{2n}$ là mẫu thuộc của tổng thể thứ hai.

Theorem

Giả sử hai tổng thể cần nghiên cứu tuân theo phân phối chuẩn với trung bình lần lượt là μ_1, μ_2. Khi đó

$$\frac{\overline{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim t(n-1)$$

Mẫu theo đôi

Chọn mẫu ngẫu nhiên gồm n cặp (X_{11}, X_{21}), $(X_{12}, X_{22}), \ldots,$ (X_{1n}, X_{2n}), trong đó $X_{11}, X_{12}, \ldots, X_{1n}$ là mẫu của tổng thể thứ nhất; $X_{21}, X_{22}, \ldots, X_{2n}$ là mẫu thuộc của tổng thể thứ hai.
Đặt $D_i = X_{1i} - X_{2i}, i = 1, \ldots, n$ là những đại lượng chỉ sự sai lệch trên từng cặp.

Theorem

Giả sử hai tổng thể cần nghiên cứu tuân theo phân phối chuẩn với trung bình lần lượt là μ_1, μ_2. Khi đó

$$\frac{\overline{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim t(n-1)$$

Mẫu theo đôi

Chọn mẫu ngẫu nhiên gồm n cặp (X_{11}, X_{21}), $(X_{12}, X_{22}), \ldots,$ (X_{1n}, X_{2n}), trong đó $X_{11}, X_{12}, \ldots, X_{1n}$ là mẫu của tổng thể thứ nhất; $X_{21}, X_{22}, \ldots, X_{2n}$ là mẫu thuộc của tổng thể thứ hai.
Đặt $D_i = X_{1i} - X_{2i}, i = 1, \ldots, n$ là những đại lượng chỉ sự sai lệch trên từng cặp.
Gọi D là đại lượng nhận giá trị $D_1, D_2, \ldots, D_n$ thì khi đó D là mẫu ngẫu nhiên các sai lệch trên mỗi cặp. Do đó ta có thể lập $\overline{D}$ là trung bình của mẫu D và S_D^2 là phương sai mẫu ngẫu nhiên D.

Mẫu theo đôi

Chọn mẫu ngẫu nhiên gồm n cặp (X_{11}, X_{21}), $(X_{12}, X_{22}), \ldots,$ (X_{1n}, X_{2n}), trong đó $X_{11}, X_{12}, \ldots, X_{1n}$ là mẫu của tổng thể thứ nhất; $X_{21}, X_{22}, \ldots, X_{2n}$ là mẫu thuộc của tổng thể thứ hai.
Đặt $D_i = X_{1i} - X_{2i}, i = 1, \ldots, n$ là những đại lượng chỉ sự sai lệch trên từng cặp.
Gọi D là đại lượng nhận giá trị $D_1, D_2, \ldots, D_n$ thì khi đó D là mẫu ngẫu nhiên các sai lệch trên mỗi cặp. Do đó ta có thể lập $\overline{D}$ là trung bình của mẫu D và S_D^2 là phương sai mẫu ngẫu nhiên D.

Theorem

Giả sử hai tổng thể cần nghiên cứu tuân theo phân phối chuẩn với trung bình lần lượt là μ_1, μ_2. Khi đó

$$\frac{\overline{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \sim t(n-1)$$

Quy luật bác bỏ và chấp nhận H_0

Quy luật bác bỏ và chấp nhận H_0

H_0	H_1	Thống kê t	Qui luật bác bỏ H_0	p-giá trị
$\mu_1 - \mu_2 \leq D_0$	$\mu_1 - \mu_2 > D_0$	$\dfrac{\overline{D} - D_0}{S_D/\sqrt{n}}$	$t > t_{n-1,\alpha}$	$P(t_{n-1} > t)$
$\mu_1 - \mu_2 \geq D_0$	$\mu_1 - \mu_2 < D_0$		$t < -t_{n-1,\alpha}$	$P(t_{n-1} < t)$
$\mu_1 - \mu_2 = D_0$	$\mu_1 - \mu_2 \neq D_0$		$\lvert t\rvert > t_{n-1,\alpha/2}$	$2P(t_{n-1} > \lvert t\rvert)$

Example

Để đánh giá một chương trình xóa đói giảm nghèo ở một vùng miền núi người ta chọn ngẫu nhiên 20 xã và thống kê tỉ lệ hộ nghèo thời điểm trước khi tiến hành chương trình, sau vài năm hoàn thành chương trình họ lại đến 20 xã trên và thống kê lại tỉ lệ hộ nghèo (%), số liệu cho bởi bảng dưới đây.
Giả sử tỉ lệ hộ nghèo của tổng thể các xã tuân theo phân phối chuẩn.
Kiểm định ở mức ý nghĩa 5% rằng trung bình tỉ lệ hộ nghèo giảm ít nhất 3% sau khi thực hiên chương trình trên.

T	8	10	11	10	10	12	8	11	10	14	9	9	8	9	9	7	8	17	8	12
S	8	8	5	6	7	6	8	3	6	8	7	8	10	8	5	2	6	5	9	8

Solution

Gọi μ_1, μ_2 lần lượt là tỉ lệ hộ nghèo trung bình của tổng thể các xã trước và sau khi thực hiện chương trình trên.

- *Cặp giả thuyết:* $H_0 : \mu_1 - \mu_2 \geq 3 \qquad H_1 : \mu_1 - \mu_2 < 3.$
- *Ta xét hiệu*

Trước	8	10	11	10	10	12	8	11	10	14	9	9	8	9	9	7	8	17	8
Sau	8	8	5	6	7	6	8	3	6	8	7	8	10	8	5	2	6	5	9
Hiệu d	0	2	6	4	3	6	0	8	4	6	2	1	-2	1	4	5	2	12	-1

- Ta tính được: $\overline{d} = 3.35, s_d = 3.313$. Do đó $t = 0.4724$.
- P - giá trị $= P(t_{19} < 0.4724) = pt(0.4724, 19) = 0.679 > 0.05$ nên chấp nhận H_0.
- Vậy, ở mức ý nghĩa 5%, chương trình xóa đói giảm nghèo trên giúp giảm tỉ lệ hộ nghèo trung bình xuống ít nhất là 3%.

Trong R

Đối với kiểm định dùng phân phối t, mẫu theo đôi, ta dùng hàm

$$t.test(\text{Mẫu 1}, \text{Mẫu 2}, mu = D_0, alt, paired = TRUE)$$

Example

Chẳng hạn trong ví dụ trên kia, ta có thể tính P - giá trị trên R bằng dãy lệnh sau:

```
> T=scan()
1: 8 10 11 10 10 12 8 11 10 14 9 9 8 9 9 7 8 17 8
20:
Read 19 items
> S=scan()
1: 8 8 5 6 7 6 8 3 6 8 7 8 10 8 5 2 6 5 9
20:
Read 19 items
> t.test(T,S,alt="less",mu=0,paird=TRUE)

        Welch Two Sample t-test

data:  T and S
t = 4.6166, df = 34.911, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 4.529371
```

Với P - giá trị $= 1 > 0.05$, ta chấp nhận H_0.

Trong R

Kiểm định trung bình hai tổng thể:

- có phân phối chuẩn, biết phương sai tương ứng σ_1^2, σ_2^2, mẫu chọn độc lập ta dùng:

$$z.test(\text{Mẫu 1}, \text{Mẫu 2}, sigma.x = \sigma_1, sigma.y = \sigma_2, mu = D_0, alt)$$

- có phân phối chuẩn, chưa biết phương sai, mẫu chọn độc lập:

$$t.test(\text{Mẫu 1}, \text{Mẫu 2}, mu = D_0, alt, var.equal = TRUE/FALSE)$$

- mà hai cỡ mẫu lớn, mẫu chọn độc lập:

$$t.test(\text{Mẫu 1}, \text{Mẫu 2}, mu = D_0, alt)$$

- phân phối chuẩn, mẫu chọn theo đôi:

$$t.test(\text{Mẫu 1}, \text{Mẫu 2}, mu = D_0, alt, paired = TRUE)$$

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

Example

Trong file "ChiTieu2010.csv" là một số thông tin về chi tiêu của một mẫu những hộ gia đình ở miền bắc, đơn vị tiền (chi tiêu) ở đây là nghìn đồng. Tại mức ý nghĩa 5%:

1. Kiểm định khẳng định cho rằng trung bình các hộ diện nghèo tiêu cho y tế ít hơn các hộ không nghèo.
2. Kiểm định nhận định rằng các hộ gia đình dùng nhiều tiền hơn cho giáo dục so với cho chăm sóc y tế.
3. có thể khẳng định được rằng trung bình chi tiêu cho ăn uống nhiều hơn 2000 (nghìn) so với chi tiêu ngoài ăn uống mỗi tháng không?
4. có thể khẳng định chi tiêu ngoại trú và nội trú là như nhau không?

Example (Ôn tập)

Dữ liệu ChiTieu2010.csv là mẫu điều tra ngẫu nhiên vài chục nghìn hộ gia đình ở nước ta. Từ đó, tại mức ý nghĩa 5% hãy thực hiện các kiểm định sau

1. Kiểm định khẳng định cho rằng trung bình một năm các hộ gia đình nước ta dành cho chi tiêu điều nội trú nhiều hơn chi tiêu điều trị ngoại trú. Mẫu được chọn là theo đôi hay độc lập?
2. Kiểm định khẳng định cho rằng chi giáo dục trung bình của các hộ ở khu thành thị (khu vực 1) là cao hơn so với nông thôn (khu vực 2). Mẫu được chọn là độc lập hay theo đôi?

Trong các lời giải đó giải thích vì sao lại dùng kiểm định t hay kiểm định z.

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

So sánh tỉ lệ hai tổng thể

Giả sử ta có hai tổng thể 1 và 2. Gọi p_1, p_2 lần lượt là tỉ lệ có dấu hiệu mà ta đang quan tâm ở hai tổng thể. Ta có các bài toán sau:

	Bài toán bên trái	Bài toán hai bên	Bài toán bên phải
H_0	$p_1 \geq p_2$	$p_1 = p_2$	$p_1 \leq p_2$
H_1	$p_1 < p_2$	$p_1 \neq p_2$	$p_1 > p_2$

Theorem

Giả sử ta có mẫu cỡ n_1 chọn từ tổng thể 1 và mẫu cỡ n_2 chọn từ tổng thể 2. Hai mẫu này được chọn độc lập. Gọi $\hat{p}_1, \hat{p}_2$ lần lượt là tỉ lệ đối tượng có dấu hiệu đang xét trong trong hai mẫu tương ứng. Khi đó $\hat{p}_1, \hat{p}_2$ là hai biến ngẫu nhiên phụ thuộc vào mẫu chọn. Do đó đại lượng

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \tag{1}$$

là một biến ngẫu nhiên.
Nếu như cỡ mẫu là lớn và số lượng các biểu hiện là không quá lệch, tức là $5 \leq n_1\hat{p}_1 \leq n_1 - 5, 5 \leq n_2\hat{p}_2 \leq n_2 - 5$ thì đại lượng Z nói trên xấp xỉ phân phối chuẩn hóa.

Quy trình kiểm định

H_0	H_1	Giá trị thống kê z	Bác bỏ H_0	p-giá trị
$p_1 - p_2 \leq 0$	$p_1 - p_2 > 0$	$\dfrac{(\hat{p_1} - \hat{p_2})}{\sqrt{\dfrac{\hat{p_1}(1-\hat{p_1})}{n_1} + \dfrac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$	$z > z_\alpha$	$P(Z > z)$
$p_1 - p_2 \geq 0$	$p_1 - p_2 < 0$		$z < -z_\alpha$	$P(Z < z)$
$p_1 - p_2 = 0$	$p_1 - p_2 \neq 0$		$\lvert z \rvert > z_{\alpha/2}$	$2P(Z > \lvert z \rvert)$

Example

Để so sánh tỉ lệ sinh viên ra trường kiếm được việc đúng ngành đào tạo của hai trường A và B, người ta chọn ngẫu nhiên 250 sinh viên tốt nghiệp trường A thấy có 130 người đang làm việc đúng chuyên ngành được đào tạo; đối với trường B người ta chọn 300 sinh viên đã tốt nghiệp và thấy có 145 sinh viên đang có công việc đúng chuyên ngành được đào tạo. Hỏi với mức ý nghĩa 5% liệu có thể nói rằng tỉ lệ sinh viên ra trường có việc làm đúng chuyên ngành được đào tạo của trường A có thấp hơn trường B không?

Solution

Gọi p_A, p_B lần lượt là tỉ lệ sinh viên trường A, B ra trường xin được việc đúng ngành đào tạo.

1. *Cặp giả thuyết:* $H_0 : p_A - p_B \geq 0 \qquad H_1 : p_A - p_B < 0.$
2. *Giá trị kiểm định:* $\hat{p_1} = 130/250, \hat{p_2} = 145/300$ *thay vào biểu thức giá trị kiểm định trong bảng trên ta được:* $z \approx 0.8569394.$
3. *Tính* $P - value = P(Z < 0.8569394) = 0.8042608.$ *Ta có P - giá trị lớn hơn 5% nên chấp nhận* $H_0.$
4. *Vậy, tại mức ý nghĩa 5%, tỉ lệ xin được việc đúng chuyên ngành đào tạo của trường A là không cao hơn so với tỉ lệ đó của trường B.*

Câu hỏi

Khảo sát cho thấy trong số 400 sinh viên học ngành kế toán ra trường có 300 sinh viên có việc làm, trong khi đó chỉ có 200 sinh viên học quản trị kinh doanh trong số 300 sinh viên được khảo sát đang có việc làm. Tại mức ý nghĩa 5%, có thể cho rằng tỉ lệ xin được việc của sinh viên học ngành kế toán là cao hơn so với tỉ lệ đó của tổng thể sinh viên học ngành quản trị kinh doanh?

Trong R

Kiểm định so sánh tỉ lệ biểu hiện T trong hai tổng thể:

prop.test(x,n,correct,alt)

Trong đó,

- x = c(số biểu hiện T trong mẫu 1, số biểu hiện T trong mẫu 2)
- n = c(cỡ mẫu 1, cỡ mẫu 2)
- correct = FALSE nếu 5 $\leq$ số biểu hiện T trong mỗi mẫu $\leq$ cỡ mẫu - 5.

Example

Trong ví dụ trên, $x = c(130, 145), n = c(250, 300)$ thỏa mãn:
$5 \leq 130 \leq 250 - 5, 5 \leq 250 \leq 300 - 5$

```
> x=c(130,145);n=c(250,300)
> prop.test(x,n,alternative = "less",correct = FALSE)

        2-sample test for equality of proportions without continuity c

data:  x out of n
X-squared = 0.73333, df = 1, p-value = 0.8041
alternative hypothesis: less
95 percent confidence interval:
 -1.0000000  0.1070466
sample estimates:
   prop 1    prop 2
0.5200000 0.4833333
```

Với p - giá trị $= 0.8041 > 0.05$ nên ta chấp nhận H_0.

Example

Từ dữ liệu ChiTieu2010.csv, hãy kiểm định tại mức ý nghĩa 5% cho khẳng định tỉ lệ chi tiêu cho ăn uống hàng tháng trên 1000 ở tổng thể hộ không nghèo là cao hơn so với tổng thể hộ nghèo.

Solution

Gọi P_1, P_2 lần lượt là tỉ lệ chi tiêu ăn uống hàng tháng trên 1000 của tổng thể các hộ không nghèo và của tổng thể các hộ nghèo.

$$H_0 : P_1 - P_2 \leq 0 \qquad H_1 : P_1 - P_2 > 0$$

```
> table(HoNgheo,CTAnUongTrongThang>1000)

HoNgheo FALSE TRUE
      0   411 6878
      1  1162  947
> table(HoNgheo)
HoNgheo
   0    1
7289 2109
```

Ta có $5 \leq 6878 \leq 7289 - 5,5 \leq 947 \leq 2109$.

Solution

```
> x=c(6878,947);n=c(7289,2109)
> prop.test(x,n,alternative = "g",correct = FALSE)

        2-sample test for equality of proportions withou

data:  x out of n
X-squared = 2871.1, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.4762246 1.0000000
sample estimates:
   prop 1    prop 2
0.9436137 0.4490280
```

P - giá trị $< 2.2 \times 10^{-16} < 0.05$ nên bác bỏ H_0.
Vậy, với xác suất sai không quá 5%, ta có thể cho rằng tỉ lệ chi tiêu cho ăn uống trên 1000 ở tổng thể hộ không nghèo là cao hơn so với tỉ lệ đó ở hộ nghèo.

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

So sánh phương sai hai tổng thể

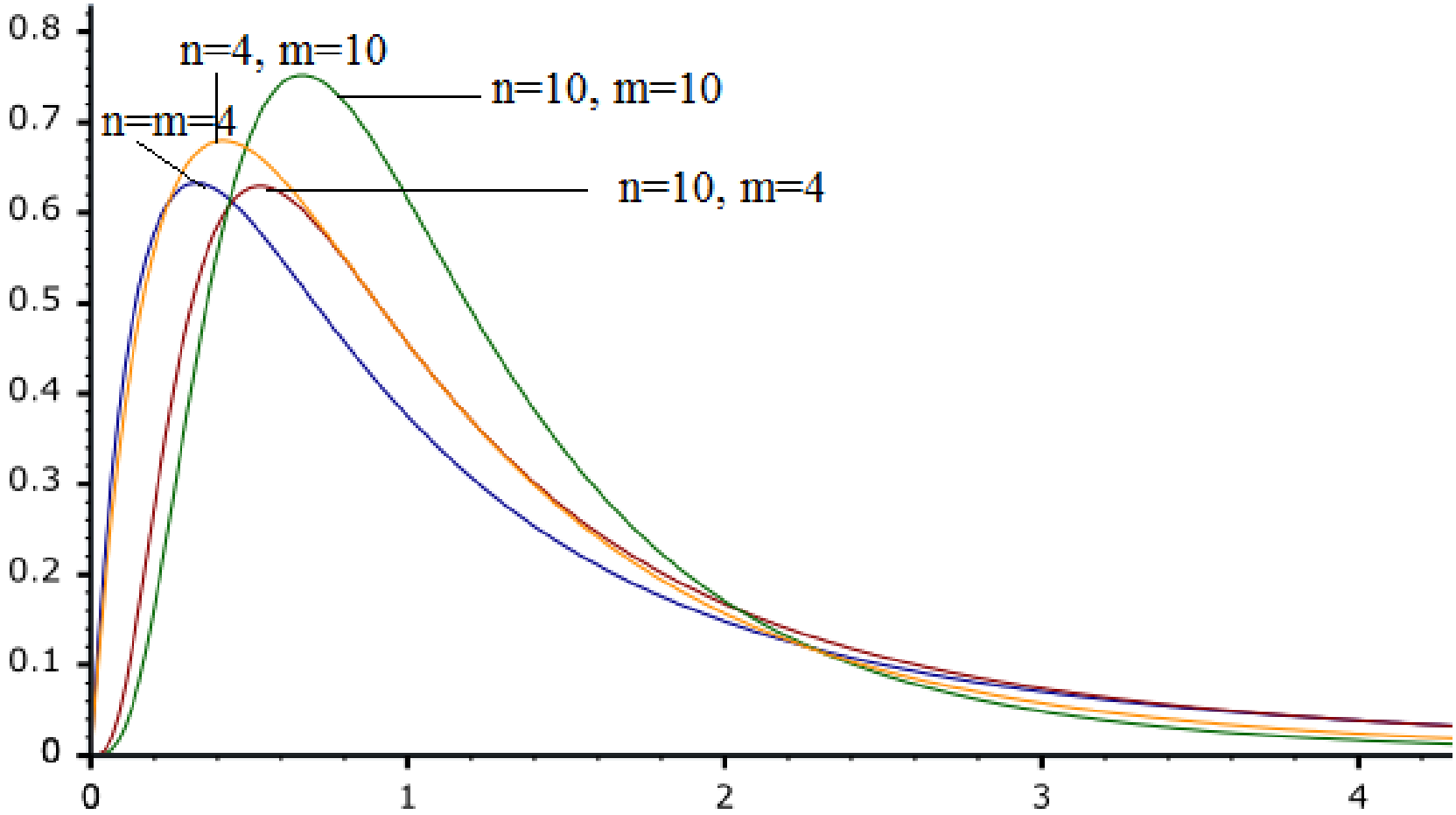
Giả sử ta có hai tổng thể 1 và 2. Gọi σ_1^2, σ_2^2 lần lượt là phương sai của hai tổng thể. Ta có các bài toán sau:

	Bài toán bên trái	Bài toán hai bên	Bài toán bên phải
H_0	$\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \leq \sigma_2^2$
H_1	$\sigma_1^2 < \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$

Theorem

Giả sử từ mỗi tổng thể chọn ra hai mẫu ngẫu nhiên có cỡ là n_1, n_2, với phương sai mẫu là S_1^2, S_2^2. Khi đó nếu hai tổng thể tuân theo phân phối chuẩn thì biến ngẫu nhiên $F = \dfrac{S_1^2}{S_2^2}$ có phân phối Fisher với $n_1 - 1$ bậc tự do ở tử và $n_2 - 1$ bậc tự do ở mẫu.

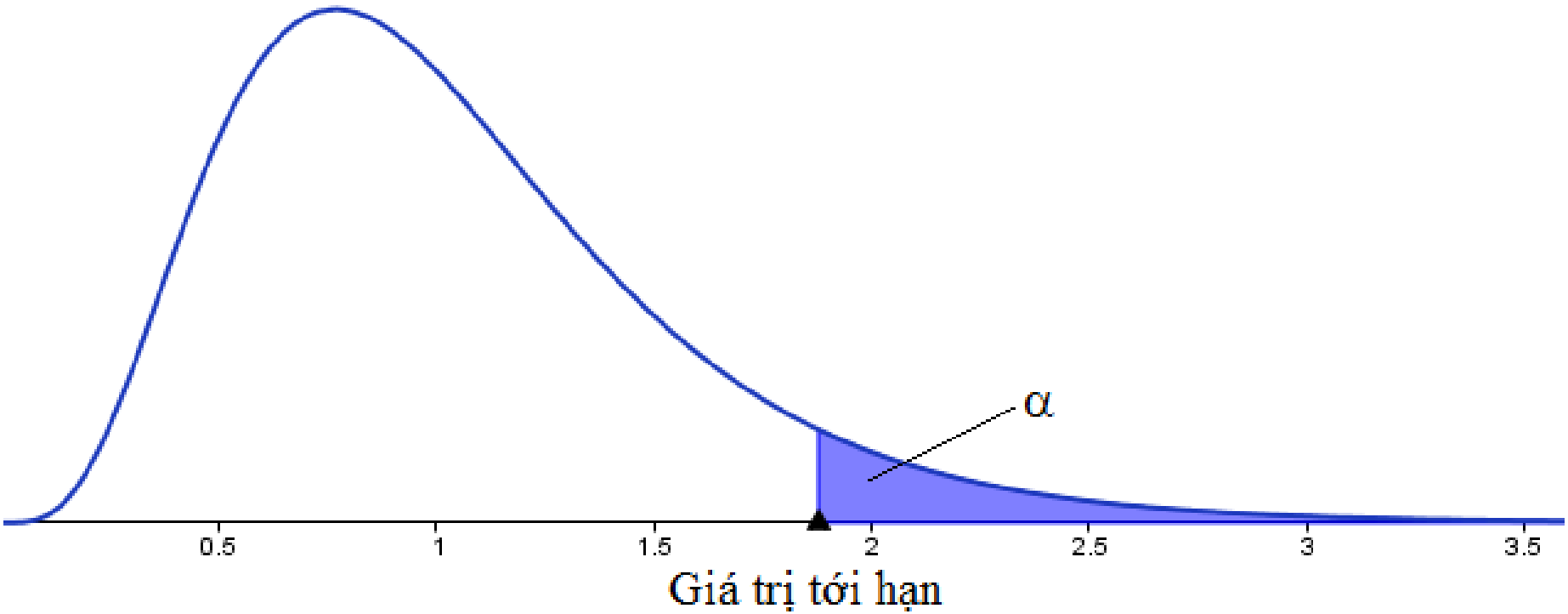
Phân phối Fisher



Giá trị tới hạn phải

Với mỗi cặp bậc tự do ta xác định giá trị $F_{m,n,\alpha}$ cho bởi:

$$P(F_{m,n} > F_{m,n,\alpha}) = \alpha$$

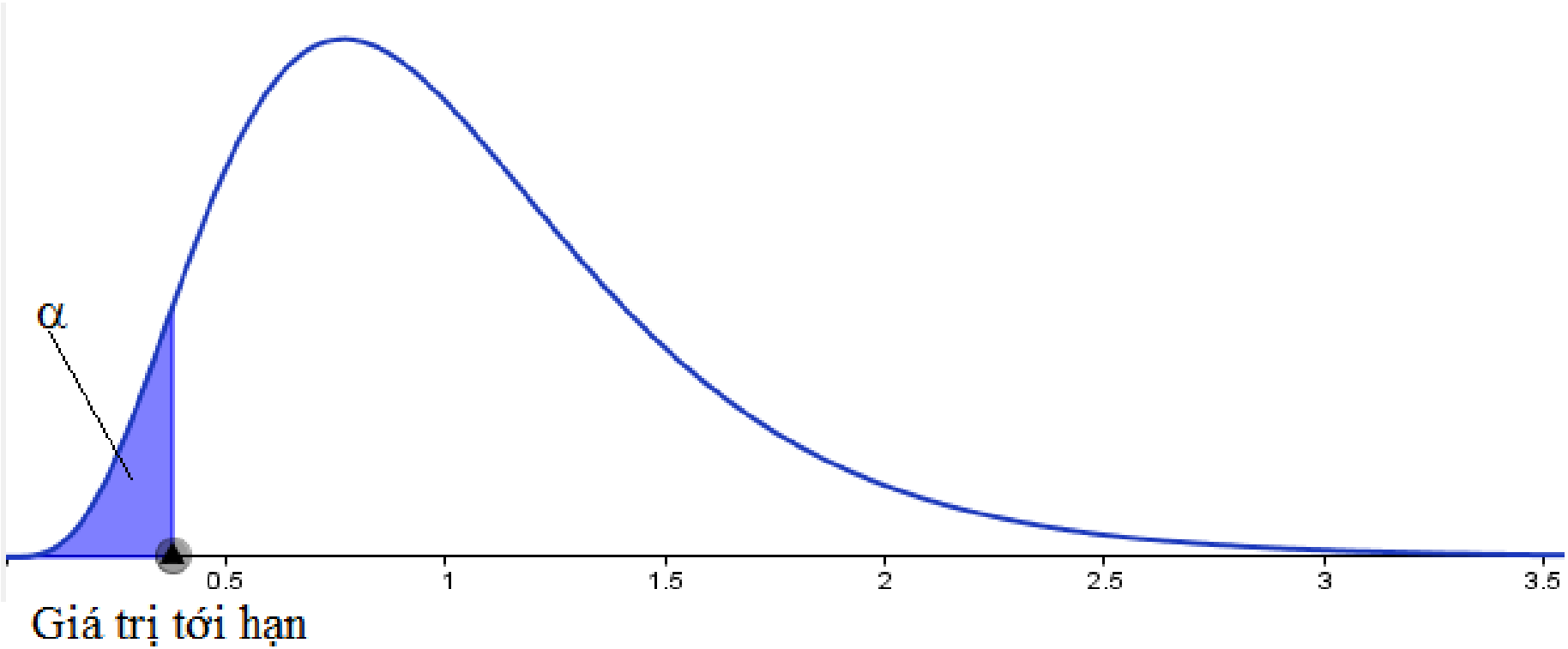


Hình: Giá trị tới hạn bên phải: $F_{m,n,\alpha}$

Giá trị tới hạn trái

Với mỗi cặp bậc tự do ta xác định giá trị $F_{m,n,1-\alpha}$ cho bởi:

$$P(F_{m,n} < F_{m,n,1-\alpha}) = \alpha$$



Hình: Giá trị tới hạn bên trái: $F_{m,n,1-\alpha}$

Kiểm định dựa trên giá trị tới hạn

H_0	H_1	Giá trị thống kê F	Qui luật bác bỏ H_0
1 $\sigma_1^2 - \sigma_2^2 \leq 0$	$\sigma_1^2 - \sigma_2^2 > 0$		$F > F_{n_1-1,n_2-1,\alpha}$
$\sigma_1^2 - \sigma_2^2 \geq 0$	$\sigma_1^2 - \sigma_2^2 < 0$	$F = \frac{s_1^2}{s_2^2}$	$F < F_{n_1-1,n_2-1,1-\alpha}$
$\sigma_1^2 - \sigma_2^2 = 0$	$\sigma_1^2 - \sigma_2^2 = 0$		$\begin{bmatrix} F > F_{n_1-1,n_2-1,\alpha/2} \\ F < F_{n_1-1,n_2-1,1-\alpha/2} \end{bmatrix}$

Lưu ý hai giá trị $F_{n_1-1,n_2-1,\alpha/2}$, $F_{n_1-1,n_2-1,1-\alpha/2}$ là nghịch đảo của nhau và trong R ta có:

$$F_{n_1-1,n_2-1,\alpha} = qf(1-\alpha, n_1-1, n_2-1)$$

Kiểm định dựa trên P - giá trị

H_0	H_1	Thống kê F	P - giá trị
$\sigma_1^2 - \sigma_2^2 \leq 0$	$\sigma_1^2 - \sigma_2^2 > 0$		$P(F_{n_1-1,n_2-1} > F)$
$\sigma_1^2 - \sigma_2^2 \geq 0$	$\sigma_1^2 - \sigma_2^2 < 0$	$F = \frac{s_1^2}{s_2^2}$	$P(F_{n_1-1,n_2-1} < F)$
$\sigma_1^2 - \sigma_2^2 = 0$	$\sigma_1^2 - \sigma_2^2 = 0$		$2.P(F_{n_1-1,n_2-1} > \max\{F, \frac{1}{F}\})$

Trong R:

$$P(F_{n_1-1,n_2-1} < x_0) = pf(x_0, n_1 - 1, n_2 - 1)$$

$$P(F_{n_1-1,n_2-1} > x_0) = pf(x_0, n_1 - 1, n_2 - 1, lower.tail = FALSE)$$

Example

Một thử nghiệm so sánh phương pháp học tập độc lập và phương pháp học nhóm được tiến hành ở hai lớp có mức học tương đương cho thấy. Nhóm học độc lập, với 100 sinh viên được kết quả trung bình 7.5 độ lệch chuẩn 1.5; nhóm còn lại với 120 sinh viên đạt trung bình 7.6 và độ lệch chuẩn là 1.2. Giả sử tổng thể điểm của sinh viên theo hai phương pháp đều có phân phối chuẩn. Tại mức ý nghĩa 5%, mức độ đồng đều về điểm số của hai tổng thể ứng với hai phương pháp học có như nhau không?

Solution

Gọi σ_1, σ_2 lần lượt là độ lệch chuẩn của tổng thể điểm của các sinh viên học theo phương pháp độc lập và học nhóm.

1. $H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2.$
2. *Nếu H_0 đúng, đại lượng* $F = \dfrac{S_1^2}{S_2^2} \sim F(99, 119).$

   Trên mẫu đã xét ta có $F = \dfrac{1.5^2}{1.2^2} = 1.5625.$
3. *Ta tính được P - giá trị* $= 2P(\ F_{99,119} > 1.5625) =$ *2*pf*$((1.5/1.2)^2$*,99,119,lower.tail* $= F) =$ *0.01998317*
4. *So sánh ta có P - giá trị* < 0.05 *nên bác bỏ* H_0.
5. *Vậy, tại mức ý nghĩa 5%, sự đồng đều về điểm của các sinh viên ở hai tổng thể tương ứng với hai phương pháp trên là khác nhau.*

Câu hỏi

Người ta muốn so sánh chỉ số IQ của những đứa trẻ hay chơi cờ với những đứa trẻ hay chơi game. Họ điều chọn được 15 cặp sinh đôi, trong mỗi cặp có 1 bé ham chơi game, 1 bé ham chơi cờ. Ta giả định rằng hai tổng thể có phân bố chuẩn. Trước khi so sánh trung bình, người ta phải xem nó có được coi là có phương sai như nhau hay không. Dựa vào mẫu sau đây, hãy trả lời câu hỏi đó ở mức ý nghĩa 5%.

Cặp	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Chơi game	*126*	*115*	*133*	*136*	*111*	*89*	*101*	*126*	*110*	*122*
Chơi cờ	*117*	*138*	*111*	*148*	*106*	*119*	*125*	*120*	*134*	*109*

Trong R

Trong bài toán kiểm định so sánh tỉ lệ phương sai của hai tổng thể $\frac{V_1}{V_2}$ với một số A, ta dùng hàm:

$$var.test(\text{Mẫu 1}, \text{Mẫu 2}, alt, ratio = A)$$

Example

Từ dữ liệu trong file "ChiTieu2010.csv", kiểm định khẳng định cho rằng phương sai trong chi tiêu cho ăn uống hàng tháng của hai tổng thể hộ nghèo và tổng thể hộ không nghèo là như nhau. Giả sử rằng hai tổng thể trên đều có phân bố chuẩn.

Solution

Gọi V_1, V_2 lần lượt là phương sai chi tiêu ăn uống hàng tháng của tổng thể các hộ nghèo và tổng thể các hộ không nghèo.

$$H_0 : V_1/V_2 = 1 \qquad H_1 : V_1/V_2 \neq 1$$

Solution

```
> x=CTAnUongTrongThang[HoNgheo==1]
> y=CTAnUongTrongThang[HoNgheo==0]
> var.test(x,y,alternative = "t",ratio = 1)

        F test to compare two variances

data:  x and y
F = 0.071749, num df = 2108, denom df = 7288, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.06703860 0.07689138
sample estimates:
ratio of variances
        0.07174904
```

Ta có p - giá trị $< 2.2 \times 0^{-16} < 0.05$ *nên bác bỏ* H_0.
Vậy, với xác suất sai lầm không quá 5%, ta có thể cho rằng phương sai của hai tổng thể nói trên là khác nhau.

1. Kiểm định giả thuyết về sự khác biệt của trung bình hai tổng thể
   - Kiểm định trung bình hai tổng thể, tổng thể chuẩn và biết phương sai, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chưa rõ phân phối, mẫu độc lập cỡ lớn
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai chưa biết và không bằng nhau, mẫu độc lập
   - Kiểm định trung bình hai tổng thể chuẩn, phương sai bằng nhau nhưng chưa biết, mẫu độc lập
   - Kiểm định trung bình hai tổng thể khi chọn mẫu theo đôi
   - Kiểm định trung bình hai tổng thể với R
   - Các ví dụ thực hành
2. Kiểm định giả thuyết cho tỉ lệ hai tổng thể
3. Kiểm định giả thuyết cho phương sai của hai tổng thể
   - Kiểm định so sánh tỉ lệ và phương sai hai tổng thể trong R

So sánh:

- tỉ lệ hai tổng thể:

$$prop.test(c(x, y), c(n_1, n_2), alt, correct = TRUE/FALSE)$$

ở đó x, y là số phần tử có dấu hiệu đang xét trong mẫu 1, 2 tương ứng và n_1, n_2 là cỡ mẫu tương ứng. Còn correct = FALSE nếu $5 \leq x \leq n_1 - 5, 5 \leq < y < n_2 - 5$; TRUE trong các trường hợp còn lại.

- phương sai hai tổng thể với giả thiết hai tổng thể có phân phối chuẩn:

$$var.test(\text{Mẫu 1}, \text{Mẫu 2}, alt, ratio = 1)$$

trong R, mặc định $ratio = 1$, nên không nhất thiết điền trong hàm.

Câu hỏi

Từ dữ liệu ChiTieu2010.csv, hãy kiểm định những khẳng định sau tại mức ý nghĩa 5%:

1. *Tỉ lệ hộ nghèo ở nông thôn là cao hơn thành thị.*
2. *Phương sai của chi tiêu giáo dục của tổng thể hộ gia đình ở nông thôn và của của tổng thể các hộ gia đình ở thành thị là ngang nhau. Giả sử chi tiêu cho giáo dục của hai tổng thể đều có phân bố chuẩn.*