Basics of Hypothesis Testing

October 2, 2023

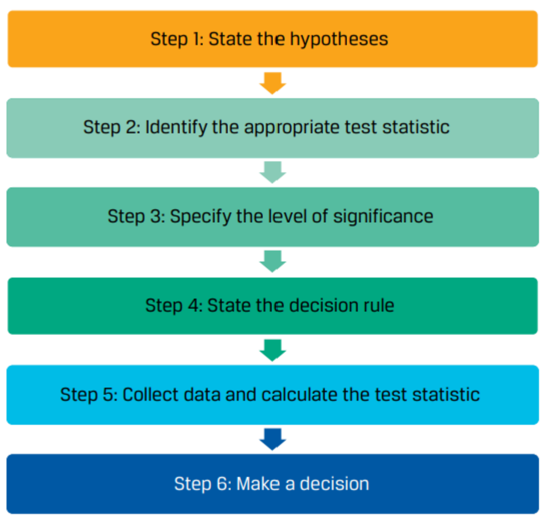
[ ]:
```
from IPython.display import Image
```

1 The Process of Hypothesis Testing

Một giả thuyết là một tuyên bố về một hoặc nhiều tổng thể cần được kiểm định thông qua thống kê mẫu (`sample statistics`)

[ ]:
```
# Quy trình kiểm định giả thuyết
Image(filename = "Pictures/01.png")
```

[ ]:



1.1 State the hypotheses

Với mỗi kiểm định giả thuyết, phân tích viên luôn luôn phải xác định cặp giả thuyết: giả thuyết "không" H_0 (`null hypothesis`) và giả thuyết thay thế H_a (`alternative hypothesis`)

Ba dạng kiểm định giả thuyết:

- Kiểm định hai phía: $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$
- Kiểm định bên phải: $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$
- Kiểm định bên trái: $H_0 : \mu \geq \mu_0$ vs $H_a : \mu < \mu_0$

Note *Giả thuyết không được kỳ vọng bác bỏ (`hoping to reject`)*

1.2 Identifying the Appropriate Test Statistic

Giá trị thống kê kiểm định (`test statistic`) là một giá trị được tính toán dựa trên cơ sở mẫu. Giá trị thử nghiệm và quy tắc quyết định (`decision rule`) là cơ sở của việc liệu giả thuyết không có bị bác bỏ hay không

Giá trị thống kê kiểm định là một biến ngẫu nhiên, vì thế phân tích viên cần quan tâm đến phân phối của nó

[ ]:
```
# Một số kiểm định tiêu biểu
Image(filename = "Pictures/02.png")
```

[ ]:

What We Want to Test	Test Statistic	Probability Distribution of the Statistic	Degrees of Freedom
Test of a single mean	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	*t*-Distributed	$n - 1$
Test of the difference in means	$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$	*t*-Distributed	$n_1 + n_2 - 2$
Test of the mean of differences	$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$	*t*-Distributed	$n - 1$
Test of a single variance	$\chi^2 = \frac{s^2(n-1)}{\sigma_0^2}$	Chi-square distributed	$n - 1$
Test of the difference in variances	$F = \frac{s_1^2}{s_2^2}$	*F*-distributed	$n_1 - 1, n_2 - 1$
Test of a correlation	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	*t*-Distributed	$n - 2$
Test of independence (categorical data)	$\chi^2 = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	Chi-square distributed	$(r-1)(c-1)$

Note: μ_0, μ_{d0}, and σ_0^2 denote hypothesized values of the mean, mean difference, and variance, respectively. The $\bar{x}$, $\bar{d}$, s^2, s, and r denote for a sample the mean, mean of the differences, variance, standard deviation, and correlation, respectively, with subscripts indicating the sample, if appropriate. The sample size is indicated as n, and the subscript indicates the sample, if appropriate. O_{ij} and E_{ij} are observed and expected frequencies, respectively, with r indicating the number of rows and c indicating the number of columns in the contingency table.

1.3 Specify the Level of Significance

Mức ý nghĩa (`level of significance`) cho biết độ mạnh tối thiểu của bằng chứng thống kê để bác bỏ giả thuyết không. Tiêu chuẩn của bằng chứng thống kê có thể thay đổi tủy vào bản chất của giả thuyết cũng như mức độ nghiêm trọng của hậu quả khi phạm sai lầm. Có 4 kết quả có thể có khi kiểm định giả thuyết không, trong đó có 2 loại sai lầm:

- Sai lầm loại I: Bác bỏ H_0 mặc dù thực tế H_0 đúng
- Sai lầm loại II: Không bác bỏ H_0 mặc dù thực tế H_0 sai

Note: *Trong nhiều trường hợp, sai lầm loại II thường được cho là nghiêm trọng hơn nhiều so với sai lầm loại I*

[ ]:

```
# Sai lầm loại I và Sai lầm loại II
Image(filename = "Pictures/03.png")
```

[ ]:

Decision	True Situation: H_0 True	True Situation: H_0 False
Fail to reject H_0	Correct decision: Do not reject a true null hypothesis.	Type II error: Fail to reject a false null hypothesis. False negative
Reject H_0	Type I error: Reject a true null hypothesis. False positive	Correct decision: Reject a false null hypothesis.

Xác suất xảy ra sai lầm loại I trong kiểm định được gọi là *mức ý nghĩa* (`level of significance`) α, trong khi $(1-\alpha)$ được định nghĩa là *độ tin cậy* (`confidence level`) của kiểm định. Ví dụ, một kiểm định có ý nghĩa ở mức 5% có nghĩa là xác suất bác bỏ giả thuyết không (trong trường hợp giả thuyết này đúng) là 5%

Trong khi đó, sức mạnh kiểm định (`power of the test`) là xác suất bác bỏ đúng giả thuyết không - nghĩa là việc bác bỏ giả thuyết không khi nó thực sự sai. Trên thực tế, chỉ số sức mạnh kiểm định $(1-\beta)$ cũng mô tả cho xác suất xảy ra sai lầm loại II trong kiểm định giả thuyết β

[ ]:

```
# Mức ý nghĩa và sức mạnh kiểm định
Image(filename = "Pictures/04.png")
```

[ ]:

Decision	True Situation: H_0 True	True Situation: H_0 False
Fail to reject H_0	Confidence level $(1 - \alpha)$	β
Reject H_0	Level of significance α	Power of the test $(1 - \beta)$

1.4 State the Decision Rule

Trước khi ra quyết định, chúng ta cần thiết lập một quy tắc quyết định, cụ thể là trả lời cho câu hỏi: *"Khi nào chúng ta bác bỏ giả thuyết không?"*. Thông thường, chúng ta sử dụng một trong ba phương pháp:

- Phương pháp giá trị tới hạn: Xác định các giá trị tới hạn (`critical value`) ở đuôi phân phối ứng với mức ý nghĩa của kiểm định. Giả thuyết không bị bác bỏ khi và chỉ khi giá trị thống kê kiểm định nằm bên ngoài khoảng giá trị bị giới hạn bởi các giá trị giới hạn
- Phương pháp sử dụng khoảng tin cậy: Xác định khoảng tin cậy dựa trên mức ý nghĩa của kiểm định, và bác bỏ giả thuyết không nếu ước lượng nằm ngoài khoảng tin cậy
- Phương pháp p-value: *(trình bày trong phần sau)*

Tiếp theo, chúng ta thực hiện thu thập dữ liệu và tính toán các giá trị kiểm định. Chất lượng của quyết định phụ thuộc vào các yếu tố: (1) sự phù hợp của mô hình ước lượng; (2) dữ liệu có chất lượng tốt (không tồn tại thiên kiến lấy mẫu, dữ liệu đã được làm sạch, dữ liệu không tồn tại lỗi đo lường...)

1.5 Make a Decision

Ra quyết định thống kê: Đưa ra kết luận về ý nghĩa thống kê của kiểm định dựa trên quy tắc quyết định đã được đưa ra từ trước

Ra quyết định kinh tế: Dựa trên các phân tích định lượng cũng như đánh giá các vấn đề kinh tế liên quan, thực thể đưa ra quyết định kinh tế

Đôi khi, ý nghĩa thống kê không đi cùng với ý nghĩa kinh tế. Vấn đề đó có thể đến từ: kích thước mẫu nhỏ, thiếu biến, hoặc đơn giản là sự ngẫu nhiên...

2 The Role of p - value

Định nghĩa: Giá trị p là mức ý nghĩa thấp nhất mà tại đó giả thuyết không bị từ chối

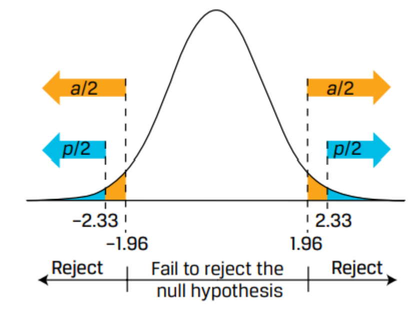
Về mặt mô phỏng, giá trị p chính là tổng diện tích bên dưới đường hàm mật độ nằm bên ngoài (các) giá trị kiểm định

Về mặt ra quyết định, giả thuyết không bị bác bỏ nếu $p < \alpha$

Đối với các phần mềm thống kê, p sinh ra mặc định cho kiểm định hai phía, vì vậy cần phải hiệu chỉnh là p trước khi ra quyết định

[ ]:
```
# Quy tắc p-value với kiểm định hai phía
Image(filename = "Pictures/05.png")
```

[ ]:



3 Multiple Tests and Significance Interpretation

Sai lầm loại I mô tả việc bác bỏ giả thuyết không trong trường hợp nó đúng (dương tính giả - `false positive`). Tỷ lệ phát hiện sai (`false discovery rate`) là đại lượng cho thấy tỷ lệ kết quả dương tính giả dự kiến. Lấy ví dụ, nếu chúng ta thực hiện 10,000 kiểm tra giả thuyết riêng biệt với mức ý nghĩa $\alpha = 5\%$, chúng ta sẽ nhận được khoảng 500 kết quả dương tính giả. Đây chính là **vấn đề kiểm định nhiều lần** (`multiple testing problem`)

Phương pháp phát hiện sai (`false discovery approach`) cho kiểm định yêu cầu hiệu chỉnh p-value nếu chúng ta có một chuỗi các kiểm định. Ý tưởng hiệu chỉnh này được giới thiệu lần đầu bởi **Benjamini & Hochberg (1995)**. Họ đề xuất xếp hạng các giá trị p từ các kiểm định và sau đó tiến hành so sánh chúng, bắt đầu từ giá trị p nhỏ nhất. Quy trình Benjamini - Hochberg gồm 4 bước:

1. Sắp xếp p-value của các kiểm định theo thứ tự tăng dần
2. Gán hạng cho p-value, giá trị nhỏ nhất nhận hạng 1, tiếp theo là hạng 2, hạng 3,...
3. Tính giá trị tới hạn Benjamini - Hochberg

$$BH(i) = \frac{i.m}{Q}$$

Trong đó, i là hạng của p-value (ở bước 2), m là tổng số kiểm định, Q là mức ý nghĩa (FDR) do phân tích viên tự lựa chọn

4. So sánh p-value và giá trị tới hạn BH, tìm p-value lớn nhất sao cho nó nhỏ hơn với giá trị tới hạn BH

Tất cả các kiểm định có p-value thấp hơn mức được xác định ở bước 4 được xác định là có ý nghĩa thống kê (bất chấp việc p-value có thể lớn hơn giá trị tới hạn BH)

Các kết luận từ việc xem xét p-value và vấn đề kiểm định nhiều lần

1. Nếu chúng ta lấy mẫu, kiểm định và phát hiện kết quả khong có ý nghĩa thống kê, kết quả này không sai; có nghĩa là giả thuyết không có thể đúng
2. Nếu sức mạnh của kiểm định (`power of the test`) thấp hoặc cỡ mẫu nhỏ; chúng ta nên thận trọng với kết quả bởi đó là cơ hội tốt để dương tính giả xuất hiện

3. Khi chúng ta thực hiện kiểm định giả thuyết và xác định giá trị tới hạn; các giá trị này dựa trên giả định rằng kiểm định được thực hiện một lần. Việc thực hiện nhiều lần một kiểm định gây ra rủi ro xuất hiện thiên kiến khai phá dữ liệu (`data snooping`) và từ đó có thể làm xuất hiện kết quả giả mạo. Hãy xác định tập dữ liệu và thực hiện kiểm định, nhưng đừng cố gắng thực hiện nhiều lần để tìm ra các kết quả có ý nghĩa thống kê, bởi vì bạn có thể tình cờ tìm thấy chúng (tức là dương tính giả)

4. Với các mẫu lớn, chúng ta hầu như tìm thấy ý nghĩa thống kê trong các kiểm định. Hướng giải quyết của vấn đề này là sử dụng các mẫu khác nhau; và nếu kết quả thu được tương tự nhau thì kết quả ấy trở nên vững hơn

4 Test Concerning a Single Mean

Các kiểm định giả thuyết liên quan đến giá trị trung bình thuộc nhóm kiểm định phổ biến nhất trong thực tế. Phân phối mẫu của giá trị trung bình là phân phối t khi không biết trước độ lệch chuẩn tổng thể; và là phân phối z trong trường hợp còn lại

Nếu tổng thể được lấy mẫu chưa biết trước phương sai, thì giá trị kiểm định cho các giả thuyết liên quan đến trung bình tổng thể duy nhất μ là:

$$t_{n-1} = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}$$

Trong đó μ_0 là giá trị giả thuyết trung bình tổng thể

Đối với các mẫu lớn, phân tích viên đôi khi sử dụng kiểm định z thay cho kiểm định t cho các kiểm định liên quan đến giá trị trung bình, bởi (1) Giá trị trung bình mẫu trong các mẫu lớn tuân theo phân phối tiệm cận chuẩn; (2) Sự khác biệt giữa kiểm định t và kiểm định z là không đáng kể với bậc tự do lớn.

Một số ít trường hợp, chúng ta có thể biết trước phương sai tổng thể. Khi đó, kiểm định z là chính xác về mặt lý thuyết:

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

[ ]:
```
# Giá trị tới hạn của kiểm định z
Image(filename = "Pictures/06.png")
```

[ ]:

Level of Significance	Alternative	Reject the Null if . . . below the Critical Value	Reject the Null if . . . above the Critical Value
0.01	Two sided: H_0: μ = $μ_0$, H_a: μ ≠ $μ_0$	−2.576	2.576
	One sided: H_0: μ ≤ $μ_0$, H_a: μ > $μ_0$		2.326
	One sided: H_0: μ ≥ $μ_0$, H_a: μ < $μ_0$	−2.326	
0.05	Two sided: H_0: μ = $μ_0$, H_a: μ ≠ $μ_0$	−1.960	1.960
	One sided: H_0: μ ≤ $μ_0$, H_a: μ > $μ_0$		1.645
	One sided: H_0: μ ≥ $μ_0$, H_a: μ < $μ_0$	−1.645	

5 Test Concerning Differences between Means with Independent Samples

Trong một số trường hợp, chúng ta muốn kiểm tra xem chênh lệch giá trị trung bình của hai tổng thể có thỏa mãn điều kiện nào không.

Với giả định cả hai tổng thể đều thỏa mãn phân phối chuẩn và chưa biết trước phương sai, chúng ta sử dụng kiểm định t dựa trên hai mẫu ngẫu nhiên độc lập:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_p^2}{n_1} + \dfrac{s_p^2}{n_2}}}$$

Trong đó:

- s_p^2 là ước lượng gộp của phương sai thông thường. Bản chất của s_p^2 là trung bình trọng số của các phương sai mẫu

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Bậc tự do của phân phối t là $n_1 + n_2 - 2$

6 Test Concerning Differences between Means with Dependent Samples

Khi chúng ta so sánh hai mẫu độc lập, chúng ta sử dụng kiểm định t với các giá trị trung bình và phương sai gộp. Kiểm định này có hiệu lực khi chúng ta giả định rằng các mẫu này độc lập với nhau. Đây gọi là kiểm định khác biệt trung bình

Tuy nhiên, trên thực tế, các mẫu có thể có liên hệ với nhau. Ví dụ như việc chúng ta cần so sánh xem liệu chính sách thuế có ảnh hưởng đến tỷ lệ chi trả cổ tức của các doanh nghiệp hay không. Khi đó, chúng ta có hai mẫu: tỷ lệ chi trả cổ tức trước và sau khi thay đổi chính sách, và rõ ràng chúng có mối liên hệ theo cặp với nhau ứng với từng công ty trong mẫu. Nhà phân tích khi này có thể kiểm định khác biệt trung bình giữa hai mẫu thông qua kiểm định so sánh theo cặp (`paired comparisons test`). Tất nhiên, chúng ta có thể sử dụng kiểm định t nếu giả định chênh lệch theo cặp của hai mẫu có phân phối chuẩn

Dưới đây là giả thuyết đối của kiểm định:

1. Kiểm định một phía: $H_a : \mu_d > \mu_{d_0}$ hoặc $H_a : \mu_d < \mu_{d_0}$
2. Kiểm định hai phía: $H_a : \mu_d \neq \mu_{d_0}$

Quy trình thực hiện kiểm định bao gồm 3 bước:

1. Xác định giá trị trung bình chênh lệch phương sai chênh lệch và phương sai trung bình chênh lệch

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \overline{d})}{n-1}}$$

$$s_{\overline{d}} = \frac{s_d}{\sqrt{n}}$$

2. Tính giá trị kiểm định t_{n-1}

$$t = \frac{\overline{d} - \mu_{d_0}}{s_{\overline{d}}}$$

3. Đưa ra quyết định dựa trên luật quyết định đã xác định từ trước

7 Test Concerning Tests of Variances

7.1 Tests of a Single Variance

Giả thuyết đối của kiểm định này được mô tả như sau:

1. Kiểm định một phía: $H_a : \sigma^2 > \sigma_0^2$ hoặc $H_a : \sigma^2 < \sigma_0^2$
2. Kiểm định hai phía: $H_a : \sigma^2 \neq \sigma_0^2$

Đối với kiểm định này, chúng ta sử dụng kiểm định chi-square (χ^2). Nếu mẫu có n quan sát, chúng ta tính được giá trị kiểm dịnh χ^2 với bậc tự do $n-1$ như sau:

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

Đến đây, chúng ta đưa ra quyết định với cặp giả định dựa trên luật quyết định đã đưa ra từ trước

7.2 Test Concerning the Equality of Two Varinaces (F-test)

Giả thuyết đối của kiểm định này được mô tả như sau:

1. Kiểm định một phía: $H_a : \sigma_1^2 > \sigma_2^2$ hoặc $H_a : \sigma_1^2 < \sigma_2^2$
2. Kiểm định hai phía: $H_a : \sigma_1^2 \neq \sigma_2^2$

Tất nhiên, hai giả thuyết trên có thể được viết lại như sau:

1. Kiểm định một phía: $H_a : \dfrac{\sigma_1^2}{\sigma_2^2} > 1$ hoặc $H_a : \dfrac{\sigma_1^2}{\sigma_2^2} < 1$
2. Kiểm định hai phía: $H_a : \dfrac{\sigma_1^2}{\sigma_2^2} \neq 1$

Ở đây, chúng ta sử dụng kiểm định F để đưa ra kết luận với các giả định trên. Giả sử chúng ta có hai mẫu với cỡ mẫu và độ lệch chuẩn mẫu lần lượt là n_1, s_1^2, n_2, s_2^2. Khi đó, chúng ta tính được giá trị kiểm định F với hai bậc tự do $n_1 - 1$ và $n_2 - 1$

$$F = \frac{s_1^2}{s_2^2}$$