Lý thuyết Thống kê (tiếp theo)

Kiểm định giả thuyết thống kê

Bài toán

Một hiệu trưởng của một trường THPT tại TPHCM đọc báo thấy rằng điểm trung bình của bài thi Đánh giá năng lực đợt 1 của Đại học Quốc gia TPHCM năm 2022 là 646,1 điểm. Hiệu trưởng nói rằng điểm trung bình của tất cả các học sinh của trường thi đáng giá năng lực lớn hơn 646,1. Sau đó, một phóng viên chọn ngẫu nhiên 50 học sinh của trường đã thi Đánh giá năng lực và thấy rằng điểm trung bình của nhóm học sinh này là 665. Như vậy, có đủ căn cứ để chấp nhận phát biểu của hiệu trưởng không?

Giả thuyết Thống kê: một dự đoán về một tham số của tổng thể.

Định nghĩa

1. **Giả thuyết** (null hypothesis), kí hiệu H_0, là một giả thuyết thống kê nói rằng **không có sự khác biệt** giữa một tham số và một giá trị cụ thể hoặc không có sự khác biệt giữa hai tham số.
2. **Đối thuyết** (alternative hypothesis), ký hiệu H_1, là một giả thuyết thống kê cho biết **có sự khác biệt** giữa một tham số và một giá trị cụ thể, hoặc có sự khác biệt giữa hai tham số.

Kiểm định giả thuyết Thống kê:

xét xem những thông tin có được từ một mẫu cụ thể có đủ cở sở để bác bỏ H_0 hay không?

Một số trường hợp thường gặp:

Kiểm định 2 phía: Giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu \neq \mu_0$.

Kiểm định 1 phía trái: Giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu < \mu_0$.

Kiểm định 1 phía phải: Giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu > \mu_0$.

Ví dụ:

1. Một giám đốc của một doanh nghiệp thấy rằng sau dịch Covid-19 mức lương trung bình của công nhân toàn công ty có thay đổi. Mức lương trung bình trước dịch Covid-19 là 8,2 triệu đồng/tháng.

   Ta kiểm định: Giả thuyết $H_0 : \mu = 8,2$ và đối thuyết $H_1 : \mu \neq 8,2$.

2. Một công nhân sản xuất gạch thấy rằng số lượng gạch làm ra trong 1 giờ giảm khi áp dụng quy trình sản xuất mới. Trước đây, trung bình công nhân làm được 35 viên gạch trong một giờ.

   Ta kiểm định: Giả thuyết $H_0 : \mu = 35$ và đối thuyết $H_1 : \mu < 35$.

3. Một nhân viên của một nhà hàng nói rằng thời gian trung bình khách phải chờ để được phục vụ của nhà hàng họ là không quá 10 phút.

   Ta kiểm định: Giả thuyết $H_0 : \mu = 10$ và đối thuyết $H_1 : \mu > 10$.

Các sai lầm trong kiểm định giả thuyết

- Sai lầm loại 1: H_0 đúng nhưng bác bỏ H_0
- Sai lầm loại 2: H_0 sai nhưng chấp nhận H_0

Trong thực tế, sai lầm loại 1 là nguy hiểm hơn, do đó ta thiết kế mô hình kiểm định sao cho xác suất sai lầm loại 1 bị chặn bởi một số rất nhỏ α.

Định nghĩa (Mức ý nghĩa - level of significance)

Xác suất bác bỏ H_0 **khi** H_0 **đúng** được gọi là **mức ý nghĩa** của kiểm định, và thường ký hiệu là α.

Giả sử biến ngẫu nhiên X có phân phối chuẩn $N(\mu; \sigma^2)$ trong đó μ là trung bình của tổng thể.

Bài toán 1. Ta kiểm định giả thuyết sau với mức ý nghĩa α

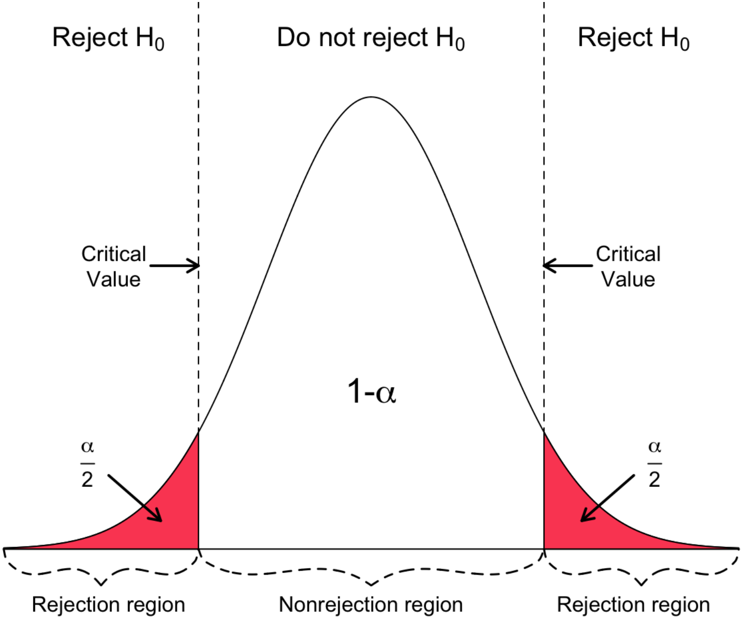
$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu \neq \mu_0.$$

Giả sử H_0 đúng. Khi đó $Z = \dfrac{\overline{X} - \mu_0}{\sigma}$ có phân phối chuẩn $N(0; 1)$.
Ký hiệu $z_{\alpha/2}$ (giá trị tới hạn) là giá trị thỏa mãn

$$P(|Z| > z_{\alpha/2}) = \alpha.$$

hay

$$\boxed{\Phi(z_{\alpha/2}) = P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2}}.$$


Reject H_0
Do not reject H_0
Reject H_0
Critical Value
Critical Value
$1-\alpha$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
Rejection region
Nonrejection region
Rejection region

Bài toán 2. Với mức ý nghĩa α, ta kiểm định:

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu > \mu_0.$$

Đặt z_α là giá trị thỏa mãn

$$P(Z > z_\alpha) = \alpha$$

hay

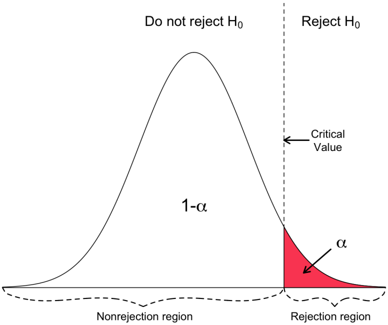
$$\boxed{\Phi(z_\alpha) = P(Z \leq z_\alpha) = 1 - \alpha}$$

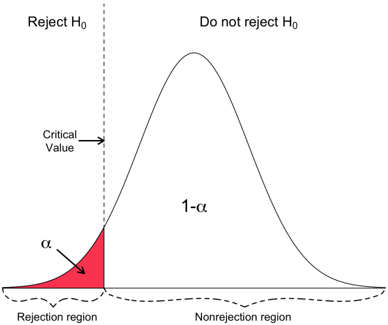
Bài toán 3. Với mức ý nghĩa α, ta kiểm định:

$$H_0 : \mu = \mu_0 \text{ và } H_1 : \mu < \mu_0.$$

Đặt z_α là giá trị thỏa mãn

$$\boxed{\Phi(z_\alpha) = P(Z < z_\alpha) = \alpha}$$


Do not reject H_0
Reject H_0
Critical Value
1-α
α
Nonrejection region
Rejection region


Reject H_0
Do not reject H_0
Critical Value
1-α
α
Rejection region
Nonrejection region

Giả sử

- $\overline{x}$ là trung bình mẫu,
- n là kích thước mẫu,
- s là độ lệch chuẩn mẫu hiệu chỉnh.

Ta xét các trường hợp sau

Trường hợp 1. σ đã biết.	**Trường hợp 2.** σ chưa biết và kích thước mẫu $n \geq 30$.
Tính $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	Tính $z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

Kiểm định	Bác bỏ H_0	Chưa bác bỏ H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$\lvert z\rvert \geq z_{\alpha/2}$	$\lvert z\rvert < z_{\alpha/2}$
$H_0 : \mu = \mu_0; H_1 : \mu > \mu_0.$	$z \geq z_\alpha$	$z < z_\alpha$
$H_0 : \mu = \mu_0; H_1 : \mu < \mu_0.$	$z \leq z_\alpha$	$z > z_\alpha$

Trường hợp 3: Với σ chưa biết và $n < 30$, tổng thể có phân phối chuẩn. Khi đó,

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim St(n)$$

trong đó $\overline{X}$ là trung bình mẫu ngẫu nhiên.

- Nếu mức ý nghĩa là α, xác định $t_{\alpha/2}$ và t_α là các số thực thỏa mãn

$$P(T > t_\alpha) = \alpha; \; P(T > t_{\alpha/2}) = \frac{\alpha}{2}.$$

- Tính thống kê mẫu

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

  trong đó $\overline{x}$ là trung bình mẫu cụ thể.

Kiểm định	Bác bỏ H_0	Chấp nhận H_0
$H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0.$	$\lvert t \rvert \geq t_{\alpha/2}$	$\lvert t \rvert < t_{\alpha/2}$
$H_0 : \mu = \mu_0; H_1 : \mu > \mu_0.$	$t \geq t_{\alpha}$	$t < t_{\alpha}$
$H_0 : \mu = \mu_0; H_1 : \mu < \mu_0.$	$t \leq -t_{\alpha}$	$t > -t_{\alpha}$

Ví dụ

Theo báo cáo "Thị trường IT Việt Nam - Developers Recruitment State 2021" do TopDev công bố cho biết, tính đến quý II/2021, kỹ sư trí tuệ nhân tạo (AI) và máy học (Machine Learning) là vị trí có mức lương trung bình hàng tháng cao nhất trong các kỹ sư IT, đạt 3054 USD (khoảng 70 triệu đồng). Một cuộc khảo sát 30 kỹ sư trí tuệ nhân tạo tốt nghiệp từ một trường đại học X cho thấy họ có mức lương trung bình là 3105 USD/tháng. Hãy kiểm tra kết luận nói rằng các kỹ sư trí tuệ nhân tạo của trường X có mức thu nhập trung bình lớn hơn 3054 USD/tháng với mức ý nghĩa 0,05. Giả sử thu nhập của các kỹ sư trí tuệ nhân tạo có phân phối chuẩn với độ lệch chuẩn tổng thể là 120 USD.

Giải.

- Gọi μ thu nhập trung bình của các kỹ sư trí tuệ nhân tạo
- Ta kiểm định: Giả thuyết $H_0 : \mu = 3054$ và đối thuyết $H_1 : \mu > 3054$
- Theo đề bài, trung bình mẫu là $\overline{x} = 3105$, cỡ mẫu $n = 30$ và độ lệch chuẩn tổng thể $\sigma = 120$
- Vì mức ý nghĩa $\alpha = 0,05$ nên $z_\alpha = 1,65$.
- Đặt
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3105 - 3054}{120/\sqrt{30}} = 2,34.$$
- Vì $z = 2,34 > 1,65$ nên bác bỏ H_0.
- Ta có căn cứ để đồng ý với tuyên bố lương trung bình của các kỹ sư trí tuệ nhân tạo nhiều hơn 3054 USD/tháng.

Ví dụ

Một nhà nghiên cứu nói rằng trung bình giá tiền của một đôi giày thể thao nam là ít hơn 80 USD. Chọn ngẫu nhiên 36 đôi giày thể thao nam để khảo sát giá, ta được kết quả sau (USD/đôi)

60	70	75	55	80	55	50	40	80
70	50	95	120	90	75	85	80	60
110	65	80	85	85	45	75	60	90
90	60	95	110	85	45	90	70	70

Giả sử giá giày có phân phối chuẩn với độ lệch chuẩn là 19,2 USD. Tuyên bố của nhà nghiên cứu có chấp nhận được không với mức ý nghĩa 10%?

Giải.

- Gọi μ giá trung bình của một đôi giày thể thao nam.
- Ta điểm định: Giả thuyết $H_0 : \mu = 80$ và đối thuyết $H_1 : \mu < 80$
- Theo đề bài, trung bình mẫu là $\overline{x} = 75$, cỡ mẫu $n = 36$ và độ lệch chuẩn tổng thể $\sigma = 19,2$
- Vì mức ý nghĩa $\alpha = 0,1$ nên $z_\alpha = -1,28$.
- Đặt
$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{75 - 80}{19,2/\sqrt{36}} = -1,56.$$
- Vì $z = -1,56 < -1,28$ nên bác bỏ H_0.
- Ta có căn cứ để đồng ý với nhận xét giá tiền trung bình của một đôi giày thể thao nam ít hơn 80 USD.

Ví dụ

Một nhà nghiên cứu nói rằng trung bình giá tiền của một đôi giày thể thao nam là ít hơn 80 USD. Chọn ngẫu nhiên 16 đôi giày thể thao nam để khảo sát giá, ta được kết quả sau (USD/đôi)

60	70	75	55	80	55	50	40
70	50	95	120	90	75	85	80

Giả sử giá giày có phân phối chuẩn. Tuyên bố của nhà nghiên cứu có chấp nhận được không với mức ý nghĩa 10%?

Giải.

- Gọi μ giá trung bình của một đôi giày thể thao nam.
- Ta điểm định: Giả thuyết $H_0 : \mu = 80$ và đối thuyết $H_1 : \mu < 80$
- Theo đề bài, trung bình mẫu là $\overline{x} = 71,875$, cỡ mẫu $n = 16$ và độ lệch chuẩn mẫu hiệu chỉnh $s =$ 20,32
- Vì mức ý nghĩa $\alpha = 0,1$ nên $t_\alpha =$ 1,341 (bậc tự do 15).
- Đặt

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{71,875 - 80}{0,32/\sqrt{16}} = -1,6$$

- Vì $t = -1,1 < 1,341$ nên Bác bỏ H_0
- ……….. ……………………………………………………

Bài toán 4

Một người ăn kiêng nói rằng có 60% số người không ăn bánh ngọt. Một cuộc khảo sát 200 người, ta thấy có 128 người nói rằng họ không ăn bánh ngọt. Với mức ý nghĩa 5%, ta có thể bác bỏ tuyên bố của người ăn kiêng này không?

- Gọi p (f, F) là tỉ lệ các phần tử có tính chất $\mathcal{P}$ trong tổng thể (trong mẫu cụ thể, mẫu ngẫu nhiên).
- Kiểm định giả thuyết

$$H_0 : p = p_0$$

- Chọn một mẫu ngẫu nhiên có kích thước n.
- Với n đủ lớn, biến ngẫu nhiên

$$Z = \frac{F - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0; 1)$$

Tính

$$z = \frac{f - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Với mức ý nghĩa α.

Kiểm định	Bác bỏ H_0	Không bác bỏ H_0
$H_0 : p = p_0; H_1 : p \neq p_0.$	$\lvert z\rvert \geq z_{\alpha/2}$	$\lvert z\rvert < z_{\alpha/2}$
$H_0 : p = p_0; H_1 : p > p_0.$	$z \geq z_\alpha$	$z < z_\alpha$
$H_0 : p = p_0; H_1 : p < p_0.$	$z \leq z_\alpha$	$z > z_\alpha$

Bài toán 4: Cho $n = 200; p_0 = 60\%$ và mức ý nghĩa 5%.

Giải Bài toán 4

- Gọi p là tỉ lệ người không ăn bánh ngọt.
- Ta kiểm định: Giả thuyết $H_0 : p = 60\%$ và đối thuyết $H_1 : p \neq 60\%$
- Tỉ lệ mẫu là $f = \dfrac{128}{200} = 0,64$ và cỡ mẫu $n = 200$.
- Vì mức ý nghĩa $\alpha = 0,05$ nên $z_{\alpha/2} = 1,96$.
- Ta có

$$z = \frac{f - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0,64 - 0,6}{\sqrt{0,6(1 - 0,6)/200}} = 1,15.$$

- Vì $|z| = 1,15 < 1,96$ nên chấp nhận H_0.
- Ta đồng ý với phát biểu rằng có 60% người không ăn bánh ngọt.

Bài tập 2

Một giáo viên nói rằng lương trung bình của giáo viên tại TPHCM ít hơn 6 triệu đồng/tháng trong năm 2021. Chọn ngẫu nhiên 8 giáo viên thì thấy mức lương hàng tháng (đơn vị là triệu đồng) của họ trong năm 2021 là

$$6 \quad 5{,}6 \quad 6 \quad 5{,}5 \quad 7 \quad 5{,}5 \quad 6 \quad 5{,}5$$

Giả sử lương của giáo viên có phân phối chuẩn. Với mức ý nghĩa 10%, tuyên bố của giáo viên đó có chấp nhận được không?

Bài tập 3

Một lập trình viên nói rằng có hơn 25% các lập trình viên đã học ngôn ngữ lập trình Python. Một cuộc khảo sát 200 lập trình viên tại một thành phố nọ, người ta thấy có 63 lập trình viên đã học Python. Với mức ý nghĩa 5%, hãy kết luận về nhận định của lập trình viên trên.

Giải Bài tập 2.

- Gọi μ tiền lương trung bình hàng tháng của giáo viên trong năm 2021.
- Ta kiểm định: Giả thuyết $H_0 : \mu = 6$ và đối thuyết $H_1 : \mu < 6$
- Trung bình mẫu là $\overline{x} = 5,88$ cỡ mẫu $n = 8$ và độ lệch chuẩn mẫu hiệu chỉnh là $s = 0,508$
- Mức ý nghĩa $\alpha = 0,1$ suy ra $t_\alpha = 1,415$ (xem bảng A5)
- Đặt

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{5,888 - 6}{0,508/\sqrt{8}} = -0,624.$$

- Vì $t = -0,624 > -1,415$ nên chấp nhận H_0.
- Ta không đồng ý với tuyên bố tiền lương trung bình trong một tháng của giáo viên ít hơn 6 triệu đồng.

Bài toán

Một cuộc khảo sát trong năm 2021 ở Việt Nam cho thấy rằng mức lương trung bình của 50 kỹ sư phần mềm là 1840 USD/tháng và 50 quản lý IT là 1750 USD/tháng. Giả sử độ lệch chuẩn tổng thể lần lượt là 400 USD và 370 USD triệu. Với mức ý nghĩa 5%, ta có thể kết luận rằng có sự khác biệt giữa mức lương trung bình của hai nhóm công việc trong ngành công nghệ thông tin không?

Ta có giả thuyết và đối thuyết lần lượt là

$$H_0 : \mu_X = \mu_Y; \; H_1 : \mu_X \neq \mu_Y.$$

với μ_X, μ_Y lần là thu nhập trung bình của kỹ sư phần mềm và quản lý IT.

Giả thuyết và đối thuyết:

$$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y \neq 0.$$

Bài toán so sánh hai trung bình của tổng thể cần các điều kiện:

- Các mẫu của hai tổng thể phải độc lập.
- Biết độ lệch chuẩn của các tổng thể
- Nếu các kích thước mẫu nhỏ hơn 30 thì các tổng thể phải có phân phối chuẩn hoặc xấp xỉ phân phối chuẩn.

Đặt $\overline{X}, \overline{Y}$ lần lượt là các trung bình của hai mẫu ngẫu nhiên được lấy từ hai tổng thể. Độ lệch chuẩn tổng thể lần lượt là σ_X, σ_Y và các kích thước mẫu lần lượt là n_X và n_Y. Khi đó

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} \sim N(0; 1).$$

Từ một bài toán so sánh trung bình hai tổng thể, ta chuyển về bài toán kiểm định trung bình của tổng thể.

Các dạng toán so sánh trung bình như sau:

So sánh	Chuyển thành
$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X \neq \mu_Y$	$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y \neq 0$
$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X < \mu_Y$	$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y < 0$
$H_0 : \mu_X = \mu_Y;\ H_1 : \mu_X > \mu_Y$	$H_0 : \mu_X - \mu_Y = 0;\ H_1 : \mu_X - \mu_Y > 0$