Lý thuyết mẫu

1. Mẫu ngẫu nhiên

Giả sử X_i là một biến ngẫu nhiên độc lập có cùng phân bố xác suất với X, tập con $(X_1, X_2, ..., X_n)$ này gọi là một **mẫu ngẫu nhiên**, n gọi là **cỡ mẫu**.

Lý thuyết mẫu

1. Mẫu ngẫu nhiên

Giả sử X_i là một biến ngẫu nhiên độc lập có cùng phân bố xác suất với X, tập con $(X_1, X_2, ..., X_n)$ này gọi là một **mẫu ngẫu nhiên**, n gọi là **cỡ mẫu**.
Khi đó kí hiệu $\mu = EX$, phương sai $\sigma^2 = DX$ thì $EX_i = \mu,\ DX_i = \sigma^2$.

Lý thuyết mẫu

1. Mẫu ngẫu nhiên

Giả sử X_i là một biến ngẫu nhiên độc lập có cùng phân bố xác suất với X, tập con $(X_1, X_2, ..., X_n)$ này gọi là một **mẫu ngẫu nhiên**, n gọi là **cỡ mẫu**.
Khi đó kí hiệu $\mu = EX$, phương sai $\sigma^2 = DX$ thì $EX_i = \mu,\ DX_i = \sigma^2$.
Gọi x_i là giá trị của X_i, khi đó $(x_1, x_2, ..., x_n)$ gọi là **giá trị của mẫu**.

2. Các đặc trưng mẫu

2.1. Trung bình mẫu

2. Các đặc trưng mẫu

2.1. Trung bình mẫu

Định nghĩa:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

2. Các đặc trưng mẫu

2.1. Trung bình mẫu

Định nghĩa:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ta có:

$$E(\overline{X}) = \mu,\ D(\overline{X}) = \frac{\sigma^2}{n}$$

2.2. Phương sai mẫu

2. Các đặc trưng mẫu

2.1. Trung bình mẫu

Định nghĩa:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ta có:

$$E(\overline{X}) = \mu,\ D(\overline{X}) = \frac{\sigma^2}{n}$$

2.2. Phương sai mẫu

Định nghĩa:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

2. Các đặc trưng mẫu

2.1. Trung bình mẫu

Định nghĩa:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ta có:

$$E(\overline{X}) = \mu,\ D(\overline{X}) = \frac{\sigma^2}{n}$$

2.2. Phương sai mẫu

Định nghĩa:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Ta có

$$E(s^2) = \sigma^2$$

2. Các đặc trưng mẫu

2.1. Trung bình mẫu

Định nghĩa:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ta có:

$$E(\overline{X}) = \mu,\ D(\overline{X}) = \frac{\sigma^2}{n}$$

2.2. Phương sai mẫu

Định nghĩa:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Ta có

$$E(s^2) = \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2$$

Ta có:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}^2$$

Ta có:

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2.$$

Khi X_i nhận các giá trị x_i thì ta có các trung bình mẫu và phương sai mẫu là các giá trị cụ thể.

Cách tính một số đặc trưng mẫu.

Trung bình mẫu

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Cách tính một số đặc trưng mẫu.

Trung bình mẫu

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Hàm trong R: **mean(x)**

Cách tính một số đặc trưng mẫu.

Trung bình mẫu

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Hàm trong R: **mean(x)**

Trung vị

Trung vị là giá trị đứng giữa của tập dữ liệu của tập dữ liệu đã được sắp thứ tự. Kí hiệu là Med và x/đ như sau:

G/s số liệu xếp theo thứ tự tăng dần thì

Med=$x_{(n+1)/2}$ nếu n lẻ.

Med=$\frac{1}{2}(x_{n/2} + x_{n/2+1})$ nếu n chẵn

Hàm trong R: median(x)

Trung vị

Trung vị là giá trị đứng giữa của tập dữ liệu của tập dữ liệu đã được sắp thứ tự. Kí hiệu là Med và x/đ như sau:

G/s số liệu xếp theo thứ tự tăng dần thì

Med=$x_{(n+1)/2}$ nếu n lẻ.

Med=$\frac{1}{2}(x_{n/2} + x_{n/2+1})$ nếu n chẵn

Hàm trong R: median(x)

Mode của dữ liệu

Là giá trị xuất hiện nhiều nhất trong bộ dữ liệu.

Mode của dữ liệu

Là giá trị xuất hiện nhiều nhất trong bộ dữ liệu.

Tứ phân vị

Tứ phân vị chia bộ dữ liệu đã sắp xếp theo thứ tự thành 4 phần có số lần xuất hiện như nhau.
Q_1: là quan sát tại vị trí 25% (n+1)
Q_2: là quan sát tại vị trí 50% (n+1)
Q_3: là quan sát tại vị trí 75% (n+1)
Hàm trong R: quantile()

Phương sai mẫu

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Hàm trong R: **var(x)**
Độ lệch tiêu chuẩn mẫu: s (sd(x)).

Khoảng biến thiên

Là hiệu số giữa giá trị lớn nhất và giá trị nhỏ nhất của dữ liệu.

Ví dụ:

Để đánh giá năng suất của một giống lúa mới người ta gieo thử trên một số mẫu ruộng và thu được kết quả sau:

Năng suất (tạ/ha)	số mẫu
25	6
30	13
33	38
34	74
35	106
36	85
37	30
39	10
40	3

Ví dụ:

Để đánh giá năng suất của một giống lúa mới người ta gieo thử trên một số mẫu ruộng và thu được kết quả sau:

Năng suất (tạ/ha)	số mẫu
25	6
30	13
33	38
34	74
35	106
36	85
37	30
39	10
40	3

Chú ý với số liệu dạng khoảng để tính các đặc trưng mẫu ta chọn điểm đại diện của mỗi khoảng

Một số định lý quan trọng về phân phối mẫu

Định lý 1

Nếu $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên quan sát phân bố chuẩn $N(\mu, \sigma^2)$ thì $\overline{X}$ cũng có phân bố chuẩn $N(\mu, \sigma^2/n)$.

Một số định lý quan trọng về phân phối mẫu

Định lý 1

Nếu $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên quan sát phân bố chuẩn $N(\mu, \sigma^2)$ thì $\overline{X}$ cũng có phân bố chuẩn $N(\mu, \sigma^2/n)$.

Định lý 2

Nếu $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên quan sát phân bố chuẩn $N(\mu, \sigma^2)$ thì $\overline{X}$ và s^2 độc lập với nhau và

$$\frac{n-1}{\sigma^2} s^2 \sim \chi^2_{n-1}$$

.

Định lý 3

Nếu $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên quan sát phân bố chuẩn thì

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

cũng có phân bố $t(n-1)$.

Định lý 3

Nếu $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên quan sát phân bố chuẩn thì

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

cũng có phân bố $t(n-1)$.

Định lý 4

Nếu $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên quan sát X bất kỳ với kỳ vọng μ, thì

$$Z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

cũng có phân bố chuẩn $N(0,1)$ khi n đủ lớn.

Định lý 5

Cho $X_1, X_2, ..., X_{n_1}$ là mẫu ngẫu nhiên được rút ra từ phân bố chuẩn $N(\mu_1, \sigma_1^2)$ và $Y_1, Y_2, ..., Y_{n_2}$ là mẫu ngẫu nhiên được rút ra từ phân bố chuẩn $N(\mu_2, \sigma_2^2)$.

Giả sử $X_1, X_2, ..., X_{n_1}$ và $Y_1, Y_2, ..., Y_{n_2}$ là hai mẫu độc lập với nhau, khi đó

$$\overline{X} - \overline{Y}$$

có phân bố chuẩn với tham số trung bình là $\mu_1 - \mu_2$, phương sai là $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Định lý 5

Cho $X_1, X_2, ..., X_{n_1}$ là mẫu ngẫu nhiên được rút ra từ phân bố chuẩn $N(\mu_1, \sigma_1^2)$ và $Y_1, Y_2, ..., Y_{n_2}$ là mẫu ngẫu nhiên được rút ra từ phân bố chuẩn $N(\mu_2, \sigma_2^2)$.
Giả sử $X_1, X_2, ..., X_{n_1}$ và $Y_1, Y_2, ..., Y_{n_2}$ là hai mẫu độc lập với nhau, khi đó

$$\overline{X} - \overline{Y}$$

có phân bố chuẩn với tham số trung bình là $\mu_1 - \mu_2$, phương sai là $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Định lý 6

Cho $X_1, X_2, ..., X_{n_1}$ là mẫu ngẫu nhiên được rút ra từ phân bố chuẩn $N(\mu_1, \sigma_1^2)$ và $Y_1, Y_2, ..., Y_{n_2}$ là mẫu ngẫu nhiên được rút ra từ phân bố chuẩn $N(\mu_2, \sigma_2^2)$. Giả sử $X_1, X_2, ..., X_{n_1}$ và $Y_1, Y_2, ..., Y_{n_2}$ là hai mẫu độc lập với nhau, khi đó tỷ số

$$F = \frac{S_1^2}{\sigma_1^2} / \frac{S_2^2}{\sigma_2^2}$$

có phân bố f_{n_1-1, n_2-1}.

Bài toán ước lượng tham số

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng.

Bài toán ước lượng tham số

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng.

Ví dụ ta muốn ước lượng trung bình của X, khi đó $\theta = EX$

Bài toán ước lượng tham số

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng.

Ví dụ ta muốn ước lượng trung bình của X, khi đó $\theta = EX$

Để tìm ước lượng cho θ ta dựa vào tập mẫu cỡ n giá trị $x_1, ..., x_n$ của X lấy ra từ tập hợp chính, ta cần tìm một giá trị θ^* xấp xỉ θ.

Bài toán ước lượng tham số

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng.

Ví dụ ta muốn ước lượng trung bình của X, khi đó $\theta = EX$

Để tìm ước lượng cho θ ta dựa vào tập mẫu cỡ n giá trị $x_1, ..., x_n$ của X lấy ra từ tập hợp chính, ta cần tìm một giá trị θ^* xấp xỉ θ.

1. Ước lượng điểm

1. Định nghĩa: **Ước lượng điểm** của θ là một hàm $\theta^* = T(X_1, ..., X_n)$ của mẫu $(X_1, ..., X_n)$.

Bài toán ước lượng tham số

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng.

Ví dụ ta muốn ước lượng trung bình của X, khi đó $\theta = EX$

Để tìm ước lượng cho θ ta dựa vào tập mẫu cỡ n giá trị $x_1, ..., x_n$ của X lấy ra từ tập hợp chính, ta cần tìm một giá trị θ^* xấp xỉ θ.

1. Ước lượng điểm

1. Định nghĩa: **Ước lượng điểm** của θ là một hàm $\theta^* = T(X_1, ..., X_n)$ của mẫu $(X_1, ..., X_n)$.
Ước lượng điểm θ^* của θ được gọi là **ước lượng không chệch** của θ nếu

$$E\theta^* = \theta$$

Bài toán ước lượng tham số

Tổng thể X chưa biết và tham số θ liên quan tới X là tham số ta cần ước lượng.

Ví dụ ta muốn ước lượng trung bình của X, khi đó $\theta = EX$

Để tìm ước lượng cho θ ta dựa vào tập mẫu cỡ n giá trị $x_1, ..., x_n$ của X lấy ra từ tập hợp chính, ta cần tìm một giá trị θ^* xấp xỉ θ.

1. Ước lượng điểm

1. Định nghĩa: **Ước lượng điểm** của θ là một hàm $\theta^* = T(X_1, ..., X_n)$ của mẫu $(X_1, ..., X_n)$.
Ước lượng điểm θ^* của θ được gọi là **ước lượng không chệch** của θ nếu

$$E\theta^* = \theta$$

Ước lượng điểm θ^* của θ được gọi là **ước lượng chệch** của θ nếu

$$E\theta^* = \theta + C$$

C được gọi là độ chệch.

Một số phương pháp tìm ước lượng

2.Phương pháp moment

Moment cấp k của X được định nghĩa như sau:

$$\mu_k = EX^k$$

Một số phương pháp tìm ước lượng

2.Phương pháp moment

Moment cấp k của X được định nghĩa như sau:

$$\mu_k = EX^k$$

Với mẫu $(X_1, X_2, ..., X_n)$ thì moment cấp k mẫu là

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$\hat{\mu}_k$ được gọi là ước lượng moment của μ.

Một số phương pháp tìm ước lượng

2.Phương pháp moment

Moment cấp k của X được định nghĩa như sau:

$$\mu_k = EX^k$$

Với mẫu $(X_1, X_2, ..., X_n)$ thì moment cấp k mẫu là

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$\hat{\mu}_k$ được gọi là ước lượng moment của μ.
Nếu $\theta_1 = f_1(\mu_1, \mu_2)$ và $\theta_2 = f_2(\mu_1, \mu_2)$ thì ước lượng moment là $\hat{\theta}_1 = f_1(\hat{\mu}_1, \hat{\mu}_2)$ và $\hat{\theta}_2 = f_2(\hat{\mu}_1, \hat{\mu}_2)$

Ví dụ:

1. Phân bố Poisson:

Moment cấp 1 của phân bố Poisson là $\lambda = EX$, do đó moment mẫu cấp 1 là

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Hay ước lượng moment cho λ là $\overline{X}$.

Ví dụ:

1. Phân bố Poisson:

Moment cấp 1 của phân bố Poisson là $\lambda = EX$, do đó moment mẫu cấp 1 là

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Hay ước lượng moment cho λ là $\overline{X}$.

2. Phân bố chuẩn: Moment cấp 1:

$$\mu_1 = EX = \mu$$

$$\mu_2 = EX^2 = \sigma^2 + \mu^2$$

Do đó:

$$\mu = \mu_1,\ \sigma^2 = \mu_2 - (\mu_1)^2$$

Như vậy các ước lượng moment từ mẫu là

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$$

2. Ước lượng hợp lý cực đại

Giả sử các biến ngẫu nhiên $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên quan sát X, θ : là tham số chưa biết cần ước lượng của X.

2. Ước lượng hợp lý cực đại

Giả sử các biến ngẫu nhiên $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên quan sát X, θ : là tham số chưa biết cần ước lượng của X.
$f(x|\theta)$ là hàm mật độ xác suất của X.

2. Ước lượng hợp lý cực đại

Giả sử các biến ngẫu nhiên $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên quan sát X, θ : là tham số chưa biết cần ước lượng của X.
$f(x|\theta)$ là hàm mật độ xác suất của X.
Các giá trị quan sát của mẫu $(x_1, x_2, ..., x_n)$.
Hàm số

$$L(\theta) = f(x_1|\theta)...f(x_n|\theta)$$

được gọi là **hàm hợp lý** của θ.

2. Ước lượng hợp lý cực đại

Giả sử các biến ngẫu nhiên $(X_1, X_2, ..., X_n)$ là mẫu ngẫu nhiên quan sát X, θ : là tham số chưa biết cần ước lượng của X.
$f(x|\theta)$ là hàm mật độ xác suất của X.
Các giá trị quan sát của mẫu $(x_1, x_2, ..., x_n)$.
Hàm số

$$L(\theta) = f(x_1|\theta)...f(x_n|\theta)$$

được gọi là **hàm hợp lý** của θ.

Định nghĩa:

Ước lượng của θ làm cực đại hàm hợp lý được gọi là **Ước lượng hợp lý cực đại.**

Hàm hợp lý

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

và loga hàm hợp lý là

$$l(\theta) = \sum_{i=1}^{n} ln(f(x_i|\theta))$$

Hàm hợp lý

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

và loga hàm hợp lý là

$$l(\theta) = \sum_{i=1}^{n} ln(f(x_i|\theta))$$

Phương trình hợp lý:

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \Longleftrightarrow \frac{\partial l(\theta)}{\partial \theta} = 0$$

Cách tìm ước lượng hợp lý cực đại:

3. Ước lượng hiệu quả và bất đẳng thức Crammer-Rao

Giả sử $\hat{\theta}$ là ước lượng của θ.

3. Ước lượng hiệu quả và bất đẳng thức Crammer-Rao

Giả sử $\hat{\theta}$ là ước lượng của θ. Khi đó sai số bình phương trung bình của $\hat{\theta}$ đối với θlà

3. Ước lượng hiệu quả và bất đẳng thức Crammer-Rao

Giả sử $\hat{\theta}$ là ước lượng của θ. Khi đó sai số bình phương trung bình của $\hat{\theta}$ đối với θlà

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = D(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$$

3. Ước lượng hiệu quả và bất đẳng thức Crammer-Rao

Giả sử $\hat{\theta}$ là ước lượng của θ. Khi đó sai số bình phương trung bình của $\hat{\theta}$ đối với θlà

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = D(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$$

Nếu ước lượng không chệch thì $MSE(\hat{\theta}) = D(\hat{\theta})$.

3. Ước lượng hiệu quả và bất đẳng thức Crammer-Rao

Giả sử $\hat{\theta}$ là ước lượng của θ. Khi đó sai số bình phương trung bình của $\hat{\theta}$ đối với θlà

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = D(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$$

Nếu ước lượng không chệch thì $MSE(\hat{\theta}) = D(\hat{\theta})$.
Tính hiệu quả của ước lượng được so sánh dưa trên phương sai của ước lượng. Giả sử $\hat{\theta}_1$ và $\hat{\theta}_2$ là hai ước lượng, khi đó

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{D(\hat{\theta}_1)}{D(\hat{\theta}_2)}$$

Ta định nghĩa lượng thông tin chứa trong mẫu $(X_1, ..., X_n)$ của tham ẩn θ là

$$I(\theta) = E\left(\frac{\partial}{\partial\theta} lnf(x|\theta)\right)^2 = -E\left(\frac{\partial^2 lnf(x|\theta)}{\partial\theta^2}\right)$$

Ta định nghĩa lượng thông tin chứa trong mẫu $(X_1, ..., X_n)$ của tham ẩn θ là

$$I(\theta) = E\left(\frac{\partial}{\partial\theta} lnf(x|\theta)\right)^2 = -E\left(\frac{\partial^2 lnf(x|\theta)}{\partial\theta^2}\right)$$

Bất đẳng thức Cramer-Rao:

Cho $X_1, X_2, ..., X_n$ là các ĐLNN độc lập cùng phân bố với hàm mật độ mẫu $f(x|\theta)$, và $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ là ước lượng không chệch của θ, khi đó ta có

$$D(\hat{\theta}) \geq \frac{1}{I(\theta)}$$

Ước lượng không chệch thỏa mãn dấu $=$ trong bất đẳng thức Cramer-Rao được gọi là ước lượng hiệu quả.

Ước lượng điểm của một số tham số quan trọng.

+ Ước lượng điểm cho trung bình là $\overline{X}$, và đó là ước lượng không chệch.

Ước lượng điểm của một số tham số quan trọng.

+ Ước lượng điểm cho trung bình là $\overline{X}$, và đó là ước lượng không chệch.

+Ước lượng điểm cho phương sai là s^2 (là ước lượng không chệch), hoặc $\hat{\sigma}^2$ (là ước lượng chệch với độ chệch là $-DX/n$.)

Ước lượng điểm của một số tham số quan trọng.

+ Ước lượng điểm cho trung bình là $\overline{X}$, và đó là ước lượng không chệch.

+Ước lượng điểm cho phương sai là s^2 (là ước lượng không chệch), hoặc $\hat{\sigma}^2$ (là ước lượng chệch với độ chệch là $-DX/n$.)

+ Ước lượng điểm cho độ lệch tiêu chuẩn là s (là ước lượng không chệch).

Ước lượng điểm của một số tham số quan trọng.

+ Ước lượng điểm cho trung bình là $\overline{X}$, và đó là ước lượng không chệch.

+Ước lượng điểm cho phương sai là s^2 (là ước lượng không chệch), hoặc $\hat{\sigma}^2$ (là ước lượng chệch với độ chệch là $-DX/n$.)

+ Ước lượng điểm cho độ lệch tiêu chuẩn là s (là ước lượng không chệch).

+Ước lượng điểm cho xác suất $p = P(A)$ là $p^* = m/n$ (m là số lần xuất hiện A trong mẫu cỡ n.

Ví dụ

Tiến hành đo chiều cao của 100 học sinh lớp 3 ở một số trường trung tiểu học ở một huyện, ta thu đươc kết quả như sau

Khoảng chiều cao (cm)	số em
$[110; 112)$	5
$[112; 114)$	8
$[114; 116)$	14
$[116; 118)$	17
$[118; 120)$	20
$[120; 122)$	16
$[122; 124)$	10
$[124; 126)$	6
$[126; 128)$	4

Ví dụ

a. Hãy ước lượng chiều cao trung bình của trẻ em lớp 3 của huyện.

b. Hãy ước lượng cho bình phương độ tản mát của chiều cao của các em học sinh lớp 3 của huyện.

c. Hãy ước lượng cho tỉ lệ học sinh có chiều cao từ 116(cm) tới 124(cm)

d. Hãy ước lượng giá trị median của chiều cao của các học sinh lớp 3 của huyện.

Ước lượng khoảng.

Định nghĩa

Ước lượng khoảng.

Định nghĩa

Một khoảng với hai đầu mút $\theta_1^* = \theta_1^*(X_1, ..., X_n)$ và $\theta_2^* = \theta_2^*(X_1, ..., X_n)$ được gọi là **ước lượng khoảng**(khoảng ước lượng, khoảng tin cậy) cho tham số θ với **độ tin cậy** (với xác suất) $1 - \alpha$ nếu

$$P(\theta_1^* < \theta < \theta_2^*) = 1 - \alpha$$

1. Ước lượng khoảng cho kỳ vọng.

Cho mẫu ngẫu nhiên $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên được rút ra từ X.

1. Ước lượng khoảng cho kỳ vọng.

Cho mẫu ngẫu nhiên $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên được rút ra từ X.

Kí hiệu: $\mu = EX$: chưa biết, $\sigma^2 = DX$.

1. Ước lượng khoảng cho kỳ vọng.

Cho mẫu ngẫu nhiên $X_1, X_2, ..., X_n$ là mẫu ngẫu nhiên được rút ra từ X.

Kí hiệu: $\mu = EX$: chưa biết, $\sigma^2 = DX$.

Công thức

Trường hợp 1: Nếu phương sai DX=σ^2 đã biết, X có phân bố chuẩn hoặc cỡ mẫu đủ lớn ($n \geq 30$) khi đó khoảng ước lượng của EX là

$$\mu \in \left(\overline{X} - z(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}; \overline{X} + z(\frac{\alpha}{2})\frac{\sigma}{\sqrt{n}}\right)$$

Trong đó $z(\alpha/2)$ được xác định theo công thức $\Phi(z(\alpha)) = 1 - \alpha$.
Hàm trong R: z.test(x, sigma.x =,conf.level =)

Trường hợp 2:

Nếu phương sai DX chưa biết, n lớn ($n \geq 30$) thì khoảng tin cậy $1 - \alpha$ của *EX* là:

Trường hợp 2:

Nếu phương sai DX chưa biết, n lớn ($n \geq 30$) thì khoảng tin cậy $1 - \alpha$ của EX là:

$$\mu \in \left(\overline{X} - z(\frac{\alpha}{2})\frac{s}{\sqrt{n}}; \overline{X} + z(\frac{\alpha}{2})\frac{s}{\sqrt{n}}\right)$$

Trường hợp 3:

Nếu phương sai DX chưa biết, n nhỏ ($n < 30$) khi đó:

Trường hợp 3:

Nếu phương sai DX chưa biết, n nhỏ ($n < 30$) khi đó:Với xác suất $1 - \alpha$, EX thuộc khoảng

$$\mu \in \left(\overline{X} - t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}; \overline{X} + t_{n-1}(\frac{\alpha}{2})\frac{s}{\sqrt{n}}\right)$$

Trong đó $t_{n-1}(\frac{\alpha}{2})$ mức phân vị của phân bố student.

Hàm trong R: t.test(x)

Ví dụ

Để nghiên cứu tuổi thọ của một dân tộc thiểu số, người ta thống kê tuổi thọ của những người đã mất của dân tộc đó trong năm qua ở các vùng miền khác nhau trên cả nước có dân tộc đó sinh sống. Kết quả như sau

Tuổi thọ (năm)	≤ 3	$(3, 10]$	$(10, 20]$	$(20, 30]$	$(30, 40]$	$(40, 50]$
Số người	15	8	4	3	2	5

$(50, 60]$	$(60, 70]$	> 70
20	18	5

a. Hãy ước lượng tuổi thọ trung bình của dân tộc đó.

b. Với độ tin cậy 95% tuổi thọ trung bình của dân tộc đó thuộc khoảng nào?

c. Với xác suất 90% có thể nói tuổi thọ trung bình của dân tộc đó cao nhất là bao nhiêu tuổi.

Ví dụ

Để xác định chiều cao trung bình của các cây bạch đàn trong khu rừng rộng trồng bạch đàn ta không đủ điều kiện đo chiều cao của mọi cây trong khu rừng, do đó người ta đo ngẫu nhiên 35 cây. Kết quả như sau

C.cao (m)	6.50 – 7.0	7.0 – 7.5	7.5 – 8.0	8.0 – 8.5	8.5 – 9.0	9.
Số cây	2	4	10	11	5	

Với xác suất 95 % ta có thể nói chiều cao trung bình của cây bạch đàn thuộc khu rừng trên nằm trong khoảng nào. Biết rằng chiều cao của các cây bạch đàn tuân theo phân bố chuẩn.

2. Ước lượng khoảng của sự khác biệt giữa 2 giá trị trung bình

TH1: G/s X, Y là 2 biến ngẫu nhiên có phân bố chuẩn với giá trị trung bình μ_1, μ_2 chưa biết, còn phương sai σ_1^2, σ_2^2 đã biết.

2. Ước lượng khoảng của sự khác biệt giữa 2 giá trị trung bình

TH1: G/s X, Y là 2 biến ngẫu nhiên có phân bố chuẩn với giá trị trung bình μ_1, μ_2 chưa biết, còn phương sai σ_1^2, σ_2^2 đã biết.

Gọi $D = \mu_1 - \mu_2 = EX - EY$.

2. Ước lượng khoảng của sự khác biệt giữa 2 giá trị trung bình

TH1: G/s X, Y là 2 biến ngẫu nhiên có phân bố chuẩn với giá trị trung bình μ_1, μ_2 chưa biết, còn phương sai σ_1^2, σ_2^2 đã biết.

Gọi $D = \mu_1 - \mu_2 = EX - EY$.

Với các mẫu ngẫu nhiên của X, Y là $(X_1, ..., X_{n_1}), (Y_1, ..., Y_{n_2})$.

2. Ước lượng khoảng của sự khác biệt giữa 2 giá trị trung bình

TH1: G/s X, Y là 2 biến ngẫu nhiên có phân bố chuẩn với giá trị trung bình μ_1, μ_2 chưa biết, còn phương sai σ_1^2, σ_2^2 đã biết.

Gọi $D = \mu_1 - \mu_2 = EX - EY$.

Với các mẫu ngẫu nhiên của X, Y là $(X_1, ..., X_{n_1}), (Y_1, ..., Y_{n_2})$.

Khi đó

$$D \in \left(\overline{X} - \overline{Y} - z(\frac{\alpha}{2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}; \overline{X} - \overline{Y} + z(\frac{\alpha}{2})\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Ví dụ

Lấy 100 quả trứng từ lô trứng do nhóm gà A đẻ ra, xác định được trọng lượng trứng trung bình là 40 gam. Lấy 120 quả từ lô trứng do nhóm gà B đẻ ra, xác định được trọng lượng trung bình là 44 gam. Với $\alpha = 5\%$, sự sai khác giữa hai loại trứng gà nằm trong khoảng nào? Biết trọng lượng quả trứng gà là tuân theo phân bố chuẩn với $\sigma_1^2 = \sigma_2^2 = 15$.

Trong trường hợp cỡ mẫu lớn thì ước lượng khoảng của $D = EX - EY$ vẫn như trên chỉ thay DX bởi s_1^2, DY bởi s_2^2.

$$D \in \left(\overline{X} - \overline{Y} - u(\frac{\alpha}{2})\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}; \overline{X} - \overline{Y} + u(\frac{\alpha}{2})\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$$

Trong trường hợp cỡ mẫu nhỏ X,Y thì khoảng ước lượng cho $EX - EY$ là:

$$D \in \left(\overline{X} - \overline{Y} - t_{n_1+n_2-2} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

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

3. Ước lượng khoảng cho tỷ lệ

Xét $p = P(A)$ chưa biết, ta cần ước lượng tỷ lệ này.

Giả sử trong mẫu cỡ n có m lần xuất hiện biến cố A, $p^* = \frac{m}{n}$ khi đó ta có:

Công thức

Với độ tin cậy $1 - \alpha$ thì tỷ lệ

$$p \in \left(p^* - z(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}, p^* + z(\frac{\alpha}{2})\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}\right)$$

Cách thực hiện trong R

library(binom)

binom.confint(x,n,method="asymptotic")

Khoảng tin cậy cho tỷ lệ này đủ tốt khi n đủ lớn và p không quá gần 0 và 1.

Ta có thể sử dụng khoảng tin cậy theo phương pháp wilson:

binom.confint(x,n,method="wilson")

Hoặc: prop.test(x,n)

Ví dụ

3. Ước lượng khoảng cho phương sai

Xét ĐLNN X với phương sai chưa biết, mẫu ngẫu nhiên quan sát X:

$X_1, X_2, ..., X_n$

3. Ước lượng khoảng cho phương sai

Xét ĐLNN X với phương sai chưa biết, mẫu ngẫu nhiên quan sát X:

$X_1, X_2, ..., X_n$

Công thức

Theo tính chất của phân phối mẫu thì $\frac{(n-1)s^2}{\sigma^2}$ có phân phối χ^2_{n-1} nên

$$P(\chi^2_{n-1}(1-\alpha/2) \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1}(\alpha/2) = 1-\alpha$$

Với độ tin cậy $1-\alpha$ thì ước lượng khoảng cho phương sai là

$$(\frac{(n-1)s^2}{\chi^2_{n-1}(\alpha/2)}, \frac{(n-1)s^2}{\chi^2_{n-1}(1-\alpha/2)})$$

Ví dụ

Kiểm tra công suất của 10 cặp pin ta thu được kết quả sau

140 136 150 144 148 152 138 141 143 151

Giả sử công suất của các cặp pin tuân theo phân phối chuẩn. Hãy

1. Cho biết một ước lượng điểm của phương sai tổng thể (tổng thể công suất tất cả các cặp pin).

2. Tìm khoảng tin cậy 99% cho phương sai tổng thể.

2. Kiểm định giả thiết

2.1. Khái niệm cơ bản

Trong chương này chúng ta đề cập đến vấn đề quan trọng của thống kê: Đó là vấn đề kiểm định giả thiết thống kê, nội dung của bài toán như sau.

2. Kiểm định giả thiết

2.1. Khái niệm cơ bản

Trong chương này chúng ta đề cập đến vấn đề quan trọng của thống kê: Đó là vấn đề kiểm định giả thiết thống kê, nội dung của bài toán như sau.
Căn cứ vào số liệu thu được (mẫu thu đươc) hãy cho một kết luận về một giả thiết thống kê nào đó mà ta đang quan tâm.

2. Kiểm định giả thiết

2.1. Khái niệm cơ bản

Trong chương này chúng ta đề cập đến vấn đề quan trọng của thống kê: Đó là vấn đề kiểm định giả thiết thống kê, nội dung của bài toán như sau.
Căn cứ vào số liệu thu được (mẫu thu đươc) hãy cho một kết luận về một giả thiết thống kê nào đó mà ta đang quan tâm.
Ví du:
Giả thiết đối chứng gọi là đối thiết.
Và vấn đề đặt ra là ta cần phải lựa chọn hoặc đối thiết(H_1) hoặc giả thiết(H_0).

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Bác bỏ H_0 thì chấp nhận H_1 và ngược lại ta chấp nhận H_0 cho đến khi có thêm thông tin mới.

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Bác bỏ H_0 thì chấp nhận H_1 và ngược lại ta chấp nhận H_0 cho đến khi có thêm thông tin mới.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên.

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Bác bỏ H_0 thì chấp nhận H_1 và ngược lại ta chấp nhận H_0 cho đến khi có thêm thông tin mới.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên.

Vận dụng các kết quả của lý thuyết xác suất ta sẽ tìm một miền S, sao cho khi mẫu $(X_1, ..., X_n) \in S$ thì ta bác bỏ giả thiết H_0, còn khi $(X_1, ..., X_n) \notin S$ thì ta chấp nhận H_0.

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Bác bỏ H_0 thì chấp nhận H_1 và ngược lại ta chấp nhận H_0 cho đến khi có thêm thông tin mới.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên.

Vận dụng các kết quả của lý thuyết xác suất ta sẽ tìm một miền S, sao cho khi mẫu $(X_1, ..., X_n) \in S$ thì ta bác bỏ giả thiết H_0, còn khi $(X_1, ..., X_n) \notin S$ thì ta chấp nhận H_0.

Khi bác bỏ hay chấp nhận H_0 ta có thể mắc phải 2 sai lầm.

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Bác bỏ H_0 thì chấp nhận H_1 và ngược lại ta chấp nhận H_0 cho đến khi có thêm thông tin mới.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên.

Vận dụng các kết quả của lý thuyết xác suất ta sẽ tìm một miền S, sao cho khi mẫu $(X_1, ..., X_n) \in S$ thì ta bác bỏ giả thiết H_0, còn khi $(X_1, ..., X_n) \notin S$ thì ta chấp nhận H_0.

Khi bác bỏ hay chấp nhận H_0 ta có thể mắc phải 2 sai lầm.

Sai lầm loại 1="Bác bỏ H_0 nhưng trên thực tế H_0 đúng"

Ta sẽ chọn 1 trong 2 giả thiết H_0 hoặc đối thiết H_1, chọn cái nào có khả năng đúng cao hơn khả năng sai thấp hơn.

Bác bỏ H_0 thì chấp nhận H_1 và ngược lại ta chấp nhận H_0 cho đến khi có thêm thông tin mới.

Để giải quyết bài toán trên, thông tin duy nhất mà chúng ta có là một mẫu ngẫu nhiên.

Vận dụng các kết quả của lý thuyết xác suất ta sẽ tìm một miền S, sao cho khi mẫu $(X_1, ..., X_n) \in S$ thì ta bác bỏ giả thiết H_0, còn khi $(X_1, ..., X_n) \notin S$ thì ta chấp nhận H_0.

Khi bác bỏ hay chấp nhận H_0 ta có thể mắc phải 2 sai lầm.

Sai lầm loại 1="Bác bỏ H_0 nhưng trên thực tế H_0 đúng"

Sai lầm loại 2="Chấp nhận H_0 nhưng thực tế H_0 sai". Ta sẽ tìm miền S sao cho

$$P(\text{Sai lầm loại 1 }) \leq \alpha, \alpha > 0 \text{ là số cho trước cố định}$$

Kiểm định sử dụng p-vale

Nếu giả thiết H_0 cần kiểm tra là đúng thì xác suất để dữ liệu phù hợp với giả thiết H_0 là bao nhiêu?

Ý nghĩa:

p-value có thể hiểu là xác suất mắc sai lầm loại 1 tối đa khi bác bỏ giả thiết với số liệu quan sát đã biết. Có thể hiểu α là mức ý nghĩa được chọn trước, còn p-value là mức ý nghĩa được tính từ số liệu.

Nếu p-value $<\alpha$ thì ta sẽ bác bỏ giả thiết H_0 và ngược lại thì ta chưa đủ thông tin để bác bỏ H_0.

Cách tính:

2.2. Kiểm định giả thiết một tổng thể

1. Kiểm định cho giá trị trung bình.
2. Kiểm định cho tỷ lệ hay cho xác suất.
3. Kiểm định phương sai.
4. Tiêu chuẩn phù hợp.

Kiểm định cho giá trị trung bình

Ta xét bài toán so sánh giá trị trung bình với một số cho trước:

Kiểm định cho giá trị trung bình

Ta xét bài toán so sánh giá trị trung bình với một số cho trước:

Giả sử $\mu = EX$ chưa biết, μ_0 là một số cho trước, mức ý nghĩa $\alpha > 0$ cho trước.

Kiểm định cho giá trị trung bình

Ta xét bài toán so sánh giá trị trung bình với một số cho trước:

Giả sử $\mu = EX$ chưa biết, μ_0 là một số cho trước, mức ý nghĩa $\alpha > 0$ cho trước.

Bài toán 1:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu \neq \mu_0$.

Kiểm định cho giá trị trung bình

Ta xét bài toán so sánh giá trị trung bình với một số cho trước:

Giả sử $\mu = EX$ chưa biết, μ_0 là một số cho trước, mức ý nghĩa $\alpha > 0$ cho trước.

Bài toán 1:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu \neq \mu_0$.

Bài toán 2:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu > \mu_0$.

Kiểm định cho giá trị trung bình

Ta xét bài toán so sánh giá trị trung bình với một số cho trước:

Giả sử $\mu = EX$ chưa biết, μ_0 là một số cho trước, mức ý nghĩa $\alpha > 0$ cho trước.

Bài toán 1:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu \neq \mu_0$.

Bài toán 2:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu > \mu_0$.

Bài toán 3:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu < \mu_0$.

Kiểm định cho giá trị trung bình

Ta xét bài toán so sánh giá trị trung bình với một số cho trước:

Giả sử $\mu = EX$ chưa biết, μ_0 là một số cho trước, mức ý nghĩa $\alpha > 0$ cho trước.

Bài toán 1:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu \neq \mu_0$.

Bài toán 2:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu > \mu_0$.

Bài toán 3:

Giả thiết H: $\mu = \mu_0$, đối thiết K: $\mu < \mu_0$.

TH1: Nếu phương sai DX=σ^2 đã biết, X phân bố chuẩn hoặc $n \geq 30$:

TH1: Nếu phương sai DX=σ^2 đã biết, X phân bố chuẩn hoặc $n \geq 30$:

Khi đó kí hiệu:

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

TH1: Nếu phương sai DX=σ^2 đã biết, X phân bố chuẩn hoặc $n \geq 30$:

Khi đó kí hiệu:

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

Khi đó miền tiêu chuẩn là:

TH1: Nếu phương sai DX=σ^2 đã biết, X phân bố chuẩn hoặc $n \geq 30$:

Khi đó kí hiệu:

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

Khi đó miền tiêu chuẩn là:

$$S_1 = \{|T| \geq z(\frac{\alpha}{2})\} \quad S_2 = \{T \geq z(\alpha)\} \quad S_3 = \{T \leq -z(\alpha)\}$$

Cách thực hiện bài toán kiểm định này:

- Từ mẫu ta tính $\overline{X}$.

- Tính Test thống kê:

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

- Tra bảng $z(\frac{\alpha}{2})$ hoặc $z(\alpha)$

- So sánh:

Nếu là BT1 thì nếu $|T| \geq z(\alpha/2)$ thì bác bỏ giả thiết chấp nhận đối thiết K.

Nếu là BT2 thì nếu $T \geq z(\alpha)$ thì bác bỏ giả thiết chấp nhận đối thiết K.

Nếu là BT3 thì nếu $T \leq -z(\alpha)$ thì bác bỏ giả thiết chấp nhận đối thiết K.

Hàm trong R: z.test hoặc zsum.test

TH2: Nếu phương sai DX chưa biết,n nhỏ, X có phân bố chuẩn,

Test thống kê:

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

$$S_1 = \{|T| \geq t_{n-1}(\frac{\alpha}{2})\} \quad S_2 = \{T \geq t_{n-1}(\alpha)\} \quad S_3 = \{T \leq -t_{n-1}(\alpha)\}$$

TH2: Nếu phương sai DX chưa biết,n nhỏ, X có phân bố chuẩn,

Test thống kê:

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

$$S_1 = \{|T| \geq t_{n-1}(\frac{\alpha}{2})\} \quad S_2 = \{T \geq t_{n-1}(\alpha)\} \quad S_3 = \{T \leq -t_{n-1}(\alpha)\}$$

- Cách thực hiện như sau:

- Hàm trong R: t.test hoặc tsum.test

TH2: Nếu phương sai DX chưa biết,n nhỏ, X có phân bố chuẩn,

Test thống kê:

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

$$S_1 = \{|T| \geq t_{n-1}(\frac{\alpha}{2})\} \quad S_2 = \{T \geq t_{n-1}(\alpha)\} \quad S_3 = \{T \leq -t_{n-1}(\alpha)\}$$

- Cách thực hiện như sau:

- Hàm trong R: t.test hoặc tsum.test

TH3: Nếu phương sai DX chưa biết, $n \geq 30$:

TH2: Nếu phương sai DX chưa biết,n nhỏ, X có phân bố chuẩn,

Test thống kê:

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

$$S_1 = \{|T| \geq t_{n-1}(\frac{\alpha}{2})\} \quad S_2 = \{T \geq t_{n-1}(\alpha)\} \quad S_3 = \{T \leq -t_{n-1}(\alpha)\}$$

- Cách thực hiện như sau:

- Hàm trong R: t.test hoặc tsum.test

TH3: Nếu phương sai DX chưa biết, $n \geq 30$:

Test thống kê:

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

$$S_1 = \{|T| \geq z(\frac{\alpha}{2})\} \quad S_2 = \{T \geq z(\alpha)\} \quad S_3 = \{T \leq -z(\alpha)\}$$

Ví dụ:

Kiểm định cho giá trị tỉ lệ

Giả sử $p = P(A)$ là tỷ lệ chưa biết; p_0 là một tỷ lệ cho trước.

$p^* = \frac{m}{n}$ là ước lượng điểm cho p.

Ta xét bài toán kiểm định giả thiết:

Glả thiết H: $p = p_0$; Đối thiết K: $p \neq p_0 (>, <)$, $\alpha > 0$ cho trước.

Test thống kê là:

$$T = \frac{m/n - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n}$$

Kiểm định cho giá trị tỉ lệ

Giả sử $p = P(A)$ là tỷ lệ chưa biết; p_0 là một tỷ lệ cho trước.

$p^* = \frac{m}{n}$ là ước lượng điểm cho p.

Ta xét bài toán kiểm định giả thiết:

Glả thiết H: $p = p_0$; Đối thiết K: $p \neq p_0(>, <)$, $\alpha > 0$ cho trước.

Test thống kê là:

$$T = \frac{m/n - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

Công thức

$$S_1 = \{|T| \geq z(\frac{\alpha}{2})\} \quad S_2 = \{T \geq z(\alpha)\} \quad S_3 = \{T \leq -z(\alpha)\}$$

Hàm trong R: prop.test()

Ví dụ

Một kho hạt giống có tỉ lệ nảy mầm xác định $p_o = 0.9$. Ngẫu nhiên có một thiết bị bị hỏng làm thay đôi điều kiện bên trong kho. Hỏi tỷ lệ nãy mầm của kho hạt giống có bị giảm xuống không (với $\alpha = 0.05$)?.
Để có thông tin về tỷ lệ nãy mầm mới của kho ta làm thí nghiệm 200 hạt thấy có 140 hạt nảy mầm.

Tiêu chuẩn phù hợp χ^2

Bài toán kiểm tra số liệu quan sát có phù hợp với các tỷ lệ

$p_i, i = 1, ...k$ ($\sum_i p_i = 1$) cho trước hay không?

Ví dụ

Theo thống kê thì trước nghị định 36CP tỷ lệ các vụ tai nạn giao thông đường bộ do người đi bộ, xe đạp, xe máy, ô tô gây ra ở thành phố Z tương ứng là $10\%, 15\%, 60\%, 15\%$. Sau 3 tháng thực hiện nghi định này, Ở thành phố Z đã xảy ra 250 vụ tai nạn giao thông đường bộ, trong đó có 40 vụ do lỗi người đi bộ, 60 vụ do lỗi người đi xe đạp, 120 vụ do lỗi người đi xe máy và 30 vụ do lỗi người đi ô tô gây ra. Với mức ý nghĩa $\alpha = 0.05$ có thể kết nói rằng sau nghi định 36CP nguyên nhân gây ra tai nạn giao thông đường bộ đã thay đổi so với trước hay không? Rút ra ý nghĩa thực tiễn gì từ kết luận nhận được.

Công thức

$$\chi^2 = \sum_{i=1}^{k} \frac{(m_i - np_i)^2}{np_i} = \frac{1}{n}\sum_{i=1}^{k} \frac{m_i^2}{p_i} - n; \quad S = \{\chi^2 \geq \chi^2_{k-1}(\alpha)\}$$

2.3. Kiểm định hai tổng thể

1. So sánh trung bình hai tổng thể.
2. So sánh hai tỷ lệ.
3. So sánh hai phương sai.

1. So sánh hai trung bình

Xét hai ĐLNN X và Y. Kí hiệu:

$\mu_1 = EX$, $\mu_2 = EY$, và $\sigma_1^2 = DX$, $\sigma_2^2 = DY$.

Mẫu $(x_1, ..., x_{n_1})$ quan sát X và mẫu $(y_1, ..., y_{n_2})$ quan sát Y.

Ta xét bài toán kiểm định giả thiết:

Giả thiết H_0: $\mu_1 = \mu_2$ $(\leq, \geq)$; Đối thiết H_1: $\mu_1 \neq \mu_2 (>, <)$, $\alpha > 0$ cho trước.

I. Với mẫu Độc lập:

Ví dụ: Trong một nghiên cứu về tác dụng của Vitamin C trong điều trị bệnh cảm lạnh, 22 người bị cảm lạnh tình nguyện được chia làm hai nhóm: nhóm 1 gồm 10 người được sử dụng một ngày 4 viên thuốc chứa 1 gram vitamin C. Nhóm 2 gồm 12 người còn lại cũng dùng thuốc bề ngoài tương tự nhưng không có vitamin C và vô hại. Việc sử dụng 4 gram vitamin C mỗi ngày có làm giảm thời gian bị cảm lạnh hay không?

Công thức

TH1. Nếu phương sai σ_1^2, σ_2^2 đã biết, X, Y phân bố chuẩn hoặc $n_1, n_2 \geq 30$

Test thống kê

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Miền tiêu chuẩn là:

$$S_1 = \{|T| \geq z(\frac{\alpha}{2})\} \quad S_2 = \{T \geq z(\alpha)\} \quad S_3 = \{T \leq -z(\alpha)\}$$

$z.test(x, y, alternative = c(''two.sided'', ''less'', ''greater''), mu = 0, sigma.x =, sigma.y =, conf.level =)$

zsum.test().

Nếu phương sai DX, DY chưa biết, X, Y chưa biết có phân bố chuẩn nhưng $n_1, n_2 \geq 30$ thì ta có thể thay σ_i bởi s_i, $i = 1, 2$.

TH2. Nếu phương sai DX, DY chưa biết ($\sigma_1^2 = \sigma_2^2$), $n_i < 30, i = 1, 2$, X, Y có phân bố chuẩn.

Test thống kê $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}}$.

Miền tiêu chuẩn:

$S_1 = \{|T| \geq t_{n_1+n_2-2}(\dfrac{\alpha}{2})\}$ $S_2 = \{T \geq t_{n_1+n_2-2}(\alpha)\}$ $S_3 = \{T \leq -T_{n_1+n_2-2}(\alpha)\}$

Hàm trong R: t.test(x, y, alternative = c("two.sided", "less","greater"),mu = , conf.level =0.95, var.equal=T...)

TH3. Nếu phương sai DX, DY chưa biết ($\sigma_1^2 \neq \sigma_2^2$), $n_i < 30, i = 1, 2$, X, Y có phân bố chuẩn.

Test thống kê

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

T có phân bố t với bậc tự do

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

Hàm trong R:

t.test(x, y, alternative = c("two.sided", "less","greater"),mu = ,

conf.level =0.95, var.equal=F...)

tsum.test()

Ví dụ

Người ta thí nghiệm hai phương pháp chăn nuôi gà khác nhau, sau một tháng, kết quả tăng trọng như sau

Phương pháp I: $n_1 = 100$ con; $\overline{X} = 1.1$ kg, $s_1^2 = 0.04$

Phương pháp II: $n_2 = 150$ con; $\overline{Y} = 1.2$ kg, $s_2^2 = 0.09$ Với mức ý nghĩa $\alpha = 0,05$ có thể kết luận phương pháp II hiệu quả hơn phương pháp I hay không?

II. Hai mẫu phụ thuộc nhau

Ví dụ: Để nghiên cứu ảnh hưởng của một loại thuốc ngủ, người ta cho 12 bệnh nhân uống thuốc thật và một lần khác uống thuốc giả, số giờ ngủ của bệnh nhân được ghi lại. Với mức ý nghĩa 5% có kết luận gì về ảnh hưởng của loại thuốc ngủ trên?

Kí hiệu

$$D = X - Y$$

Bài toán kiểm định so sánh hai giá trị trung bình sẽ được đưa về bài toán so sánh trung bình của D (μ_d) với số 0.

Thực hiện tương tự như bài toán kiểm định giá trị trung bình với một số. Mẫu $(x_i, y_i) \to d_i = x_i - y_i$

Hàm trong R:

Trường hợp phương sai chưa biết, X, Y có phân bố chuẩn ta sử dụng hàm:

t.test(x, y, alternative = c("two.sided", "less", "greater"), mu=, paired = TRUE, conf.level = 0.95, ...)

So sánh hai tỷ lệ

Ví dụ: Để so sánh tỷ lệ hạt nảy mầm của hai giống lúa người ta gieo hai mâũ thử và thấy mẫu gieo 200 hạt có 170 hạt nảy mầm, mẫu 2 gieo 150 hạt có 140 hạt nảy mầm.

Giả sử p_1, p_2 là hai tỷ lệ chưa biết, ta cần so sánh hai tỷ lệ này.

Giả thiết H_0: $p_1 = p_2$/ Đối thiết H_1: $p_1 \neq p_2 (>, <)$.

Công thức

$$Z = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}};$$

So sánh hai tỷ lệ

Ví dụ: Để so sánh tỷ lệ hạt nảy mầm của hai giống lúa người ta gieo hai mâũ thử và thấy mẫu gieo 200 hạt có 170 hạt nảy mầm, mẫu 2 gieo 150 hạt có 140 hạt nảy mầm.

Giả sử p_1, p_2 là hai tỷ lệ chưa biết, ta cần so sánh hai tỷ lệ này.

Giả thiết H_0: $p_1 = p_2$/ Đối thiết H_1: $p_1 \neq p_2 (>, <)$.

Công thức

$$Z = \frac{p_1^* - p_2^*}{\sqrt{p^*(1-p^*)(\frac{1}{n_1} + \frac{1}{n_2})}};$$

$$S_1 = \{|T| \geq z(\frac{\alpha}{2})\};\ S_2 = \{T \geq u(\alpha)\};\ S_3 = \{T \leq -z(\alpha)\}$$

Hàm trong R: prop.test(x=c(), n=c())

So sánh hai phương sai

Giả sử các ĐLNN X và Y có phân bố chuẩn với σ_1^2, σ_2^2 là các phương sai chưa biết ta thiết lập bài toán kiểm định giả thiết:

Giả thiết H_0: $\sigma_1^2 = \sigma_2^2$/ Đối thiết H_1: $\sigma_1^2 \neq \sigma_2^2 (>, <)$.

Do tính chất phân phối mẫu $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ tuân theo phân bố Fisher $(n_1 - 1, n_2 - 1)$ nên

Công thức

Test thống kê

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

So sánh hai phương sai

Giả sử các ĐLNN X và Y có phân bố chuẩn với σ_1^2, σ_2^2 là các phương sai chưa biết ta thiết lập bài toán kiểm định giả thiết:

Giả thiết H_0: $\sigma_1^2 = \sigma_2^2$/ Đối thiết H_1: $\sigma_1^2 \neq \sigma_2^2 (>, <)$.

Do tính chất phân phối mẫu $F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$ tuân theo phân bố Fisher $(n_1 - 1, n_2 - 1)$ nên

Công thức

Test thống kê

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

Miền bác bỏ giả thiết:

$$S_1 = (0, f_{n_1-1,n_2-1}(1-\alpha/2) \cup (f_{n_1-1,n_2-1}(\alpha/2), \text{inf})$$

$$S_2 = (f_{n_1-1,n_2-1}(\alpha), \text{inf}); \; S_3 = (0, f_{n_1-1,n_2-1}(1-\alpha)$$

2. Phân tích phương sai

2.1. Phân tích phương sai một nhân tố

Ví dụ Một cuộc thực nghiệm để thẩm định sự khác biệt về sản lượng của một giống lúa dùng 3 loại phân bón khác nhau. Kết quả được ghi lại (theo đơn vị tạ) trong bảng sau đây:

Bón phân loại 1	86.92; 88; 77; 84
Bón phân loại 2	92; 91; 90; 81; 93
Bón phân loại 3	75; 80; 83; 79

Hãy kiểm tra xem sản lượng trung bình của giống lúa khi bón 3 loại phân trên có như nhau hay không.

Bài toán

Giả sử A là một nhân tố nào đó (A: phân bón) với k mức tác động.

μ_i : là trung bình của mức tác động thứ i của nhân tố A, $i = 1, 2, ..., k$.

Mỗi mức tác động coi là 1 nhóm.

Bài toán KĐGT:

Giả thiết H_0:

$$\mu_1 = \mu_2 = ... = \mu_k$$

Đối thiết H_1 :

Tồn tại $i \neq j$ sao cho $\mu_i \neq \mu_j$

Mô hình

x_{ij} là kết quả quan sát thứ j của nhóm i, $j = 1, 2..., n_i$ (nhóm i có n_i quan sát).

Mô hình phân tích phương sai phát biểu rằng:

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

μ : trung bình toàn bộ quần thể; α_i: ảnh hưởng của nhóm i; và sai số ϵ_{ij}. Giả sử sai số

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

Ta có:

$$x_{ij} = \overline{x} + (\overline{x_i} - \overline{x}) + (x_{ij} - \overline{x_i})$$

Trong đó: $\overline{x} = \sum_{ij} x_{ij}/n$ là trung bình chung toàn bộ mẫu;

$\overline{x_i} = \sum_j x_{ij}/n_i$ là trung bình của nhóm i.

Ta chia tổng bình phương thành 2 phần: nhân tố và sai số. Để tiến hành phân tích phương sai ta kí hiệu:

Tổng bình phương toàn bộ mẫu:

$$SST = \sum_i \sum_j (x_{ij} - \overline{x})^2$$

Tổng bình phương vì sự khác nhau giữa các nhóm (do nhân tố)

$$SSF = \sum_i \sum_j (\overline{x_i} - \overline{x})^2 = \sum_i n_i (\overline{x_i} - \overline{x})^2$$

Tổng bình phương sai số trong từng nhóm

$$SSE = \sum_i \sum_j (x_{ij} - \overline{x_i})^2$$

Ta dễ dàng chứng minh:

$$SST = SSF + SSE$$

Trung bình bình phương (PS) do nhân tố gây ra là:

$$MSF = SSF/(k-1)$$

và trung bình bình phương sai số:

$$MSE = SSE/(n-k)$$

Tỷ số $F = \frac{MSF}{MSE}$ có phân bố Fisher nếu giả thiết H_0 đúng.

Do đó nếu $F \geq F_{k-1,n-k}(\alpha)$ thì ta bác bỏ H_0.

Ta có bảng phân tích phương sai ANOVA:

Ví dụ:

Sử dụng R

Bước 1: Nhập toàn bộ mẫu vào 1 véc tơ: x

Bước 2: Tạo nhóm: Tạo một véc tơ (g) gồm thứ tự nhóm của các quan sát, sau đó biến véc tơ thành nhân tố bằng lệnh

g<- as.factor(g)

Bước 3: aov($x \sim g$)

summary(aov($x \sim g$))

Trong trường hợp bài toán ANOVA cho kết quả các giá trị trung bình là khác nhau, ta muốn phân tích sâu hơn để xác cặp trung bình nào khác nhau ta dùng phương pháp TukeyHSD như sau:

$$TukeyHSD(aov(x \sim g))$$

$$plot(TukeyHSD(aov(x \sim g)))$$

2.2. Phân tích phương sai hai nhân tố

Phân tích phương sai hai nhân tố xem xét cùng một lúc hai yếu tố nguyên nhân ảnh hưởng đến kết quả mà ta đang nghiên cứu. Ví dụ: Năng suất bị ảnh hưởng bởi lượng phân bón và giống lúa.

Gọi A, B là hai nhân tố tác động lên kết quả nghiên cứu. A có s mức tác động, B có r mức tác động.

Bài toán kiểm định như sau:

Bài toán 1:
Giả thiết H_0^1: Trung bình do các mức tác động của nhân tố A gây nên là như nhau.
Đối thiết H_1^1: Trung bình do các mức tác động của nhân tố A gây nên là khác nhau.
Bài toán 2:
Giả thiết H_0^2: Trung bình do các mức tác động của nhân tố B gây nên là như nhau.
Đối thiết H_1^2: Trung bình do các mức tác động của nhân tố B gây nên là khác nhau.

- Mẫu quan sát:Kẻ bảng theo ô, mỗi ô có một quan sát.

- Mô hình phân tích phương sai hai nhân tố

$$x_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

Giả thiết: $\epsilon_{ij} \sim N(0, \sigma^2)$

Mẫu quan sát (x_{ij})

Kí hiệu:

$\overline{x} = \sum_{ij} x_{ij}/n$ là trung bình chung toàn bộ mẫu.

$\overline{x_{i0}} = \sum_j x_{ij}/n_i$ là trung bình của mức tác động thứ i của nhân tố A.

$\overline{x_{0j}} = \sum_i x_{ij}/n_j$ là trung bình của mức tác động thứ j của nhân tố B.

Trong phân tích phương sai hai nhân tố ta chia tổng bình phương thành 3 phần:

Tổng bình phương do các mức tác động của nhân tố A gây ra

$$SSF_A = \sum_i n_i(\overline{x_{i0}} - \overline{x})^2$$

Tổng bình phương do các mức tác động của nhân tố B gây ra

$$SSF_B = \sum_j n_j(\overline{x_{0j}} - \overline{x})^2$$

Tổng bình phương sai số trong từng nhóm

$$SSE = \sum_i \sum_j (x_{ij} - \overline{x_{i0}} - \overline{x_{0j}} + \overline{x})^2$$

Bảng phân tích phương sai:

Sự phân tán	Bậc tự do	Tổng BP	Trung bình BP	Tỷ số F
Nhân tố A	s-1	SSF_A	$MSF_A = \frac{SSF_A}{s-1}$	$F_A = \frac{MSF_A}{MSE}$
Nhân tố B	r-1	SSF_B	$MSF_B = \frac{SSF_B}{r-1}$	$F_B = \frac{MSF_B}{MSE}$
Sai số	(s-1)(r-1)	SSE	$MSE = \frac{SSE}{n-sr+2}$	
Tổng số	n-1			

Miền bác bỏ giả thiết tương tự trên.

Ví dụ:

Gộp mẫu:

```
Mau= c(5.8, 6.2, 5.4, 6.0, 5.2, 5.3, 5.4, 5.6, 6.2, 5.7, 5.5, 6.1, 6.0, 5.2,
6.4, 5.5, 5.0, 5.6, 6.2, 6.1, 5.3, 6.0, 6.6, 6.1, 5.8, 5.9, 6.0, 5.9, 6.0, 6.7,
6.5, 6.3, 6.1, 6.8, 6.4, 6.8, 6.6, 6.4, 6.2, 7.1, 7.0, 7.2, 6.2, 5.8, 6.5, 6.2,
6.4, 5.7, 6.1, 6.8, 7.1, 6.5, 7.1, 7.2, 6.7, 7.0, 7.6, 7.7, 7.8, 6.8, 7.3, 7.1,
7.2)
```

Phân nhóm:

```
PhanNhomA = rep(c(1, 2, 3), c(21, 21, 21))
PhanNhomA = as.factor(PhanNhomA)
PhanNhomB = rep(c(1, 2, 3), each = 7, length = 63)
PhanNhomB= as.factor(PhanNhomB)
anova(lm(Mau ~ PhanNhomA + PhanNhomB))
```

Mô hình phân tích phương sai có sự tương tác hai nhân tố:

$$x_{ij} = \mu + \alpha_i + \beta_j + (\alpha_i \beta_j)_{ij} + \epsilon_{ij}$$

Giả thiết: $\epsilon_{ij} \sim N(0, \sigma^2)$

Mẫu quan sát (x_{ijk}), mỗi ô có hơn 1 quan sát khi đó ta có thêm bài toán sau:

Bài toán 3:

Giả thiết H_0^3: Không có sự tác động qua lại giữa hai nhân tố A,B.

Đối thiết H_1^3: Có sự tác động qua lại giữa hai nhân tố A,B.

Bảng phân tích phương sai ANOVA:

Ta dùng hàm TukeyHSD() để thực hiện phân tích sâu Two-way ANOVA trong R.

$$TukeyHSD(aov(Mau \sim PhanNhomA + PhanNhomB))$$

Minh họa bằng hình vẽ:

$$plot(TukeyHSD(aov(Mau \sim PhanNhomA)))$$

$$plot(TukeyHSD(aov(Mau \sim PhanNhomB)))$$

plot(TukeyHSD(aov(MauGop s PhanKhoi))) Minh họa sự tương tác ta dùng hàm:

$$interaction.plot(PhanNhomA, PhanNhomB, Mau)$$

3. Kiểm định phi tham số

Kiểm định tham số:

Kiểm định phi tham số:

Ưu điểm:

- Không đòi hỏi những giả định về tham số và phân phối tổng thể.

- Dùng được cho các dữ diệu định danh và thứ bậc.

Tính toán ít phức tạp vì thường cỡ mẫu nhỏ.

Nhược điểm:

- Khả năng tìm được sự khác biệt kém hơn, khi mẫu lớn các tính toán thường tẻ nhạt kém hấp dẫn.

3.1. Tiêu chuẩn Wilcoxon

Tiêu chuẩn Wilcoxon về trung vị của một tổng thể

Khi phân bố của tổng thể nghiêng hẳn sang bên trái hoặc bên phải thì trung vị sẽ là số đo tập trung tốt hơn trung bình tổng thể. Hơn nữa khi cỡ mẫu nhỏ và tổng thể không có phân phối chuẩn.

Gọi M_d là trung vị của tổng thể.

M_0 là số cho trước.

Các bài toán kiểm định giả thiết:

Bài toán 1: Giả thiết H_0: $M_d = M_0$, $H_1 : M_d \neq M_0$.

Bài toán 2: Giả thiết H_0: $M_d = M_0$, $H_1 : M_d > M_0$.

Bài toán 3: Giả thiết H_0: $M_d = M_0$, $H_1 : M_d < M_0$.

Hạng của các phần tử trong dãy: Cho dãy $\{x_1, ..., x_n\}$các phần tử được sắp theo thứ tự tăng dần.

Hạng phần tử là trung bình cộng vị trí của phần tử đó trong dãy.

Nếu trong dãy chỉ có duy nhất 1 phần tử x_i thì $rank(x_i) = i$,...

Các bước thực hiện:

- Tính $d_i = x_i - M_0$
- Xếp hạng các phần tử $|d_i|$, bỏ qua các giá trị $d_i = 0$; n= số các phần tử $d_i \neq 0$.
- Tính tổng hạng $R^+ = \sum_{d_i>0} rank(|d_i|)$

Chọn $V = R^+$

- Tính các giá trị

$$v_{\alpha,n};\ \frac{n(n+1)}{2} - v_{\alpha,n}$$

$v_{\alpha,n}$ được tính trong R bằng hàm psignrank().

- So sánh với giá trị tới hạn (bảng so sánh).

Nếu n đủ lớn (>20) H_0 đúng thì V có phân bố xấp xỉ chuẩn với trung bình $EV = \frac{n(n+1)}{4}$ và phương sai $DV = \frac{n(n+1)(2n+1)}{24}$.

Trong R: wilcox.test(x,mu,...)

Loại bỏ nhiễu trong x để tính p-value chính xác bằng hàm: jitter(x)

Tiêu chuẩn Wilcoxon so sánh trung vị hai tổng thể

Các bài toán kiểm định giả thiết:

Bài toán 1: Giả thiết H_0: $M_x = M_y$, $H_1 : M_x \neq M_y$.

Bài toán 2: Giả thiết H_0: $M_x = M_y$, $H_1 : M_x > M_y$.

Bài toán 3: Giả thiết H_0: $M_x = M_y$, $H_1 : M_x < M_y$.

Hai mẫu độc lập (Mann Whithney)

Cách thực hiện

Giả sử mẫu của x có cỡ mẫu là m, mẫu y cỡ là n.

- Gộp chung hai mẫu và tính hạng trong mẫu chung.
- Tính tổng hạng của các của x_i là W.
- Tính các giá trị phân vị của phân bố Wilcoxon:

$$v_{\alpha,m}, \frac{m(m+n+1)}{2} - v_{\alpha,m}$$

$v_{\alpha,m}$ tính trong R bằng hàm qwilcox().

- So sánh W với giá trị phân vị (bảng so sánh).

Khi cỡ mẫu đủ lớn

Khi $m + n$ đủ lớn thì

$$z = \frac{W - m(m+n+1)/2}{\sqrt{mn(m+n+1)/12}}$$

có phân bố N(0,1).
Do đó ta có thể so sánh z với z_α hoặc $z_{\alpha/2}$

Hàm trong R: wilcox.test(x, y)

Ví dụ

Một công ty muốn đánh giá số lượng hàng hóa bán ra ở hai cửa hàng, người ta điều tra doanh số bán ra của hai cửa hàng (theo tuần) thu được số liệu sau:

Cửa hàng A: 22, 34, 52, 62, 30, 40, 64, 84, 56, 59

Cửa hàng B: 52, 71, 76, 54, 67, 83, 66, 90, 77, 84

Hai mẫu phụ thuộc nhau

- Tính độ lệch giữa hai mẫu $d_i = x_i - y_i$, bỏ qua các giá trị $d_i = 0$ và tính hạng $rank(|d_i|)$.
- Tương tự như 1 tổng thể.

wilcox.test(x, y,pair=T)

Ví dụ:

Đánh giá tác dụng của một chế độ ăn bồi dưỡng mà dấu hiệu quan sát là số hồng cầu của một số người trước và sau khi ăn ta được số liệu sau:

X_i: 32,33,32,40,36,43,42,34,43,64,67,36,39,48,55,51

Y_i: 36,41,36,41,43,39,40,42,43,50,58,36,40,48,59,55

Hỏi chế độ ăn có tác dụng làm tăng số hồng cầu hay không với 5%

So sánh nhiều giá trị trung bình khi mẫu nhỏ phân bố không tuân theo phân bố chuẩn

Bài toán KĐGT:
Giả thiết H_0: $\mu_1 = \mu_2 = ... = \mu_k$
Đối thiết H_1 : Tồn tại $i \neq j$ sao cho $\mu_i \neq \mu_j$

Các mẫu độc lập

- Gộp các mẫu lại có cỡ là n và tính hạng các phần tử trong mẫu chung.
- Tính tổng hạng riêng từng mẫu R_i.

Người ta chứng minh được rằng

$$T = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

có phân bố χ^2. Do đó:
Nếu $T \geq \chi^2_{k-1,\alpha}$.

Trong R:

kruskal.test(x,...)

Nếu x là véc tơ mẫu gộp thì:

- Tạo ra vecto thứ bậc để phân mẫu:

$$g = c(rep(c(1, .., k), rep(c(n_1, ..., n_k)))$$

- kruskal.test(x,g)

Nếu x gồm nhiều véc tơ tách rời $x_1, x_2, ..., x_k$

$$kruskal.test(list(x_1, x_2, ..., x_k))$$

Ví dụ

Người ta muốn điều tra năng suất hoạt động của tổ máy cùng làm ra một loại sản phẩm. Đếm số sản phẩm làm ra trong 1 ngày của 3 máy ta có số liệu sau:

Tổ 1: 18,22,25,18,15,22,34

Tổ 2: 20,19,23,25,32,33,23

Tổ 3: 21,22,24,33,30

Kiểm định chi bình phương

Kiểm định chính xác McNemar: so sánh tỷ lệ

Với dữ liệu theo cặp: ví dụ ta điều tra mẫu ở hai thời điểm khác nhau. Xét mẫu gồm dữ liệu về việc đánh giá hiệu quả của Thủ tướng với 1600 người được hỏi và được hỏi lại sau 6 tháng.

Bài toán kiểm định giả thiết là: H_0: Tỷ lệ hai lần điều tra là như nhau.

H_1: Tỷ lệ hai lần điều tra là khác nhau.

	Test 2 positive	Test 2 negative
Test 1 positive	n_{11}	n_{12}
Test 1 negative	n_{21}	n_{22}

Gọi π_{11} là tỷ lệ cả hai lần điều tra đều dồng ý π_{12} là tỷ lệ lần điều tra thứ nhất đồng ý lần thứ 2 không đồng ý

π_{21} là tỷ lệ lần điều tra thứ nhất không đồng ý nhưng lần thứ 2 đồng ý.

π_{22} là tỷ lệ hai lần điều tra đều phản đối.

Khi đó giả thiết tương đương với: $\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21}$ và $\pi_{21} + \pi_{22} = \pi_{12} + \pi_{22}$. Hay giả thiết tương đương với

$$\pi_{12} = \pi_{21}$$

Giả sử số quan sát tương ứng với π_{ij} là n_{ij}, $i, j = 1, 2$.

Như vậy ta có Test thống kê:

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

Hoặc ta có thể sử dụng hiệu chỉnh liên tục:

$$\chi^2 = \frac{(n_{12} - n_{21} - 1)^2}{n_{12} + n_{21}}$$

Trong R: Performance <- matrix(c(794, 86, 150, 570), nrow = 2, dimnames = list("1st Survey" = c("Approve", "Disapprove"), "2nd Survey" = c("Approve", "Disapprove")))

Performance

mcnemar.test(Performance)

Kiểm tra tính độc lập

Kiểm định Chi bình phương:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

trong đó O_i là tần số quan sát; E_i là tần số lý thuyết (mong đợi).

Bài toán kiểm định:

H_0: X và Y độc lập với nhau.

H_1 : X và Y phụ thuộc nhau.

Công thức

O_{ij}: số quan sát từ mẫu.

Cộng tổng hàng:

$$r_i = \sum_{j=1}^{s} O_{ij}$$

tổng cột:

$$c_j = \sum_{i=1}^{r} O_{ij}$$

Nếu X,Y độc lập thì

$$E_{ij} = n \times \frac{r_i}{n} \times \frac{c_i}{n} = \frac{r_i \times c_j}{n}$$

Tính

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = n(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{O_{ij}^2}{r_i c_j} - 1)$$

Do đó nếu giả thiết đúng thì χ^2 sẽ có phân bố χ^2 với bậc tự do

(r-1)(s-1) khi $n_{ij} \geq 5$ với mọi i, j.

Miền tiêu chuẩn $S = \{\chi^2 \geq \chi^2_{(r-1)(s-1)}(\alpha)\}$

Trong R

Ta dùng hàm chisq.test(A), trong đó ma trận A gồm các quan sát của hai thuộc tính X và Y.
matrix(x, nrow = m, ncol = n, byrow = FALSE)
chisq.test(A)

Ví dụ

Nghiên cứu về sự phụ thuộc giữa màu mắt và màu tóc, ta có một mẫu được điều tra ngẫu nhiên như sau

	Xanh	Nâu	Đen
Đen	11	35	50
Nâu	18	12	10
Vàng	10	9	5

Với mức ý nghĩa $\alpha = 0.1$ có thể kết luận màu mắt và màu tóc độc lập với nhau không?

So sánh nhiều tỷ lệ

Bài toán: so sánh k tỷ lệ.

Cộng tổng hàng: $r_i = \sum_{j=1}^{2} O_{ij}$, tổng cột: $c_j = \sum_{i=1}^{k} O_{ij}$

Tính $\chi^2 = n(\sum_{i=1}^{2} \sum_{j=1}^{k} \frac{O_{ij}^2}{r_i \times c_j} - 1)$

Miềm tiêu chuẩn $S = \{\chi^2 \geq \chi^2_{k-1}(\alpha)\}$

Ví dụ

Một hãng sản suất ô tô muốn tìm hiểu xem có sự phụ thuộc nào giữa giới tính của người sở hữu và kiểu dáng xe ô tô hay không. Một mẫu ngẫu nhiên gồm 2000 chủ sở hữu ô tô được chon và phân loại như sau

	I	II	III
Nam	350	270	380
Nữ	340	400	260

Với mức ý nghĩa $\alpha = 0.025$, tỷ lệ nữ dùng 3 loại xe trên có như nhau hay không?

Trong trường hợp n_{ij} nhỏ, cỡ mẫu n nhỏ ta có thể sử dụng tiêu chuẩn sau:

Kiểm định chính xác Fisher

Nhập bảng số liệu định danh:
A= matrix(x,nrow=,dimnames =list())
Kiểm định Fisher: fisher.test(A,alt=" ")
Ví dụ:

Kiểm tra phân bố chuẩn

Chia miền giá trị mẫu thành các phần $S_1, ..., S_k$ sau đó xác định $p_1 = P(X \in A_1), ..., p_k = P(X \in A_k)$ và áp dụng tiêu chuẩn χ^2 để kiểm định.

Ví dụ

Tiến hành đo chiều cao của 100 cây bạch đàn trong một khu rừng trồng bạch đàn của một lâm trường ta thu được kết quả sau:

Khoảng chiều cao (m)	số cây	Khoảng chiều cao (m)	số cây
8.275 – 8.325	1	8.625 – 8.675	17
8.325 – 8.375	2	8.675 – 8.725	12
8.375 – 8.425	4	8.725 – 8.775	9
8.425 – 8.475	5	8.775 – 8.825	7
8.475 – 8.525	8	8.825 – 8.875	6
8.525 – 8.575	10	8.875 – 8.925	0
8.575 – 8.625	18	8.925 – 8.975	1

Hãy kiểm tra giả thiết cho rằng chiều cao của các cây bạch đàn là tuân theo phân bố chuẩn ($\alpha = 0.05$)?

Khoảng $S_i(a_i, a_{i+1})$	m_i	p_i	$\frac{(m_i - np_i)^2}{np_i}$
$(-\infty, 8.425)$	7	0.0548	0.4216
$(8.425, 8.427)$	5	0.0583	0.1182
$(8.475, 8.525)$	8	0.0930	0.1817
$(8.525, 8.575)$	10	0.1295	0.6720
$(8.575, 8.625)$	18	0.1484	0.6729
$(8.625, 8.675)$	17	0.1528	0.1936
$(8.675, 8.725)$	12	0.1735	0.1365
$(8.725, 8.775)$	9	0.1004	0.1077
$(8.775, 8.825)$	7	0.0650	0.0385
$(8.825, +\infty)$	7	0.0643	0.0505
$\sum$	100		$\chi^2 = 2.5923$

Với $\alpha = 0.05$ thì $\chi^2_7(0.05) = 14.0671$, vậy $\chi^2 < \chi^2_7(0.05)$ nên ta chấp nhận giả thiết, tức biến ngẫu nhiên chiều cao của cây bạch đàn là tuân theo phân bố chuẩn với độ tin cậy là 95%.

Trong R

shapiro.test(*x*)