Apenddix 1 - Lý thuyết phân phối và kiểm định thống kê

10 May 2019 - phamdinhkhanh

Menu

Phần 1 - Thống kê

1.1. Các đại lượng thống kê

1.**Đại lượng ngẫu nhiên**: Một đại lượng được coi là ngẫu nhiên nếu giá trị của nó là kết quả của một biến cố ngẫu nhiên. Chẳng hạn phép tung đồng xu đồng chất với 2 mặt xấp ngửa là một đại lượng ngẫu nhiên vì ta không hoàn toàn biết trước khả năng đồng xu rơi vào mặt xấp hoặc ngửa. Có 2 loại đại lượng ngẫu nhiên:

- Đại lượng ngẫu nhiên liên tục: Giá trị của nó có thể rơi vào bất kì một giá trị nào nằm trong một khoảng xác định. Chẳng hạn chiều cao, cân nặng của một người có thể coi là đại lượng liên tục.
- Đại lượng ngẫu nhiên rời rạc: Giá trị của nó nằm trong một tập hợp hữu hạn các khả năng. Ví dụ như trường hợp tung đồng xu ta chỉ có thể nhận các giá trị là {0, 1} tương ứng với khả năng rơi vào mặt S(sấp) hặc mặt N (ngửa).

2.**Kì vọng**: là giá trị trung bình của một đại lượng ngẫu nhiên. Giá trị của kỳ vọng được chia thành 2 trường hợp:

- Nếu x là đại lượng ngẫu nhiên rời rạc.

$$\mathrm{E(x)} = \bar{\mathrm{x}} = \sum_{i=1}^{n} x_i p(x_i)$$

Trong đó $p(x_i)$ là xác suất xảy ra biến cố $x = x_i$. Khi khả năng xảy ra của các biến cố ngẫu nhiên rời rạc x_i là như nhau thì giá trị của kỳ vọng: $\mathrm{E}(x) = \frac{\sum_{i=1}^{n} x_i}{n}$

- Nếu x là một đại lượng ngẫu nhiên liên tục:

$$\mathrm{E(x)} = \bar{\mathrm{x}} = \int xp(x)dx$$

Một số tính chất của kì vọng:

- $\mathrm{E(ax)} = a\mathrm{E(x)}$
- $\mathrm{E(ax+by)} = a\mathrm{E(x)} + b\mathrm{E(y)}$
- Nếu x, y là 2 biến ngẫu nhiên độc lập thì $\mathrm{E(xy)} = \mathrm{E(x)E(y)}$

3.**Hiệp phương sai**: Là đại lượng đo lường mối quan hệ cùng chiều hoặc ngược chiều giữa 2 biến ngẫu nhiên. Đây là đại lượng được sử dụng nhiều trong kinh tế lượng và thống kê học để giải thích mối quan hệ tác động giữa các biến. Khi hiệp phương sai giữa 2 biến lớn hơn 0, chúng có quan hệ đồng biến và ngược lại. Hiệp phương sai chỉ được tính trên 2 chuỗi có cùng độ dài.

$$\mathrm{cov(x,\,y)} = \mathrm{E}[(\mathrm{x} - \bar{\mathrm{x}})(\mathrm{y} - \bar{\mathrm{y}})] = \frac{\sum_{i=1}^{n}(x_i - \bar{\mathrm{x}})(y_i - \bar{\mathrm{y}})}{n}$$

Gía trị của hiệp phương sai giữa 2 chuỗi số x,y được kí hiệu là $\mathrm{cov(x,y)}$ hoặc σ_{xy} và được tính bằng kì vọng của tích chéo độ lệch so với trung bình của 2 đại lượng như công thức trên.

Như vậy ta có thể rút ra các tính chất của hiệp phương sai:

- tính chất giao hoán: $\mathrm{cov(x,\,y)} = \mathrm{cov(y,\,x)}$
- tính chất tuyến tính: $\mathrm{cov(ax,\,by)} = \mathrm{ab.cov(x,\,y)}$
- Khai triển công thức hiệp phương sai ta có:
  $\mathrm{cov(x,\,y)} = \mathrm{E(xy)} - \mu_{\mathrm{x}}\mathrm{E(y)} - \mu_{\mathrm{y}}\mathrm{E(x)} + \mu_{\mathrm{x}}\mu_{\mathrm{y}}$

Trong đó $\mu_{\mathrm{x}}, \mu_{\mathrm{y}}$ lần lượt là kì vọng của x, y. Chứng minh công thức trên không khó. Xin dành cho bạn đọc.

4.**Phương sai**: Là trường hợp đặc biệt của hiệp phương sai giữa một đại lượng ngẫu nhiên với chính nó. Giá trị của phương sai luôn lớn hơn hoặc bằng 0 do bằng tổng bình phương sai số của từng mẫu so với kỳ vọng. Trong trường hợp phương sai bằng 0, đại lượng là một hằng số không biến thiên. Phương sai của một đại lượng thể hiện mức độ biến động của đại lượng đó xung quanh giá trị kỳ vọng. Nếu phương sai càng lớn, miền biến thiên của đại lượng càng cao và ngược lại.

Phương sai được kí hiệu là $\mathrm{Var}(x)$, σ_x^2 hoặc s_x^2. Công thức phương sai được tính như sau:

- Nếu x là đại lượng ngẫu nhiên rời rạc:

$$\mathrm{Var}(x) = \sum_{i=1}^{n}(x_i - \mu)^2 p(x_i)dx$$

Trong đó $\mathrm{E}(x) = \mu$. Khi các biến cố xảy ra với cùng xác suất bằng $\frac{1}{n}$, phương sai chính là trung bình $\mathrm{Var}(x) = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$

- Nếu x là đại lượng ngẫu nhiên liên tục:

$$\mathrm{Var}(x) = \int (x - \mu)^2 p(x)dx$$

Phương sai của một biến có thể được tính toán thông qua kì vọng của biến:

$$\begin{aligned}\mathrm{Var}(x) &= \mathrm{E}((x-\mu)^2) \\ &= \mathrm{E}((x^2 - 2\mu x + \mu^2)) \\ &= \mathrm{E}(x^2) - 2\mu\mathrm{E}(x) + \mathrm{E}(\mu^2) \\ &= \mathrm{E}(x^2) - 2\mu^2 + \mu^2 \\ &= \mathrm{E}(x^2) - \mu^2 \\ &= \mathrm{E}(x^2) - \mathrm{E}(x)^2\end{aligned}$$

Đây là một trong những tính chất rất thường được sử dụng trong tính toán nhanh phương sai mà bạn đọc cần nhớ. Đồng thời từ công thức trên ta cũng suy ra một bất đẳng thức quan trọng đó là kỳ vọng của bình phương luôn lớn hơn bình phương của kỳ vọng: $\mathrm{E}(x^2) \geq \mathrm{E}(x)^2$

5.**Độ lệch chuẩn**: Độ lệch chuẩn của một đại lượng có giá trị bằng căn bậc 2 của phương sai. Nó đại diện cho sai số của đại lượng so với trung bình.

$$\sigma_x = \sqrt{\mathrm{Var}(x)}$$

Trong trường hợp các biến rời rạc phân phối đều với xác suất $\frac{1}{n}$:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n}}$$

Trong thống kê chúng ta thường xác định các giá trị outliers dựa trên nguyên lý 3 sigma bằng cách xem những giá trị nằm ngoài khoảng $[\mu - 3\sigma, \mu + 3\sigma]$ như là outliers. Ta có thể xử lý outliers bằng cách đưa về đầu mút gần nhất $\mu - 3\sigma$ hoặc $\mu + 3\sigma$ hoặc loại bỏ luôn outliers.

6.**Hệ số tương quan**: Là một chỉ số có quan hệ gần gũi với hiệp phương sai. Hệ số tương quan đánh giá mối quan hệ đồng biến hay nghịch biến giữa 2 đại lượng ngẫu nhiên. Tuy nhiên khác với hiệp phương sai, hệ số tương quan cho biết thêm mối quan hệ tương quan tuyến tính giữa 2 biến là mạnh hay yếu.

Hệ số tương quan giao động trong khoảng [-1, 1]. Tại 2 giá trị đầu mút -1 và 1, hai biến hoàn toàn tương quan tuyến tính. Tức ta có thể biểu diễn $\mathrm{y} = a\mathrm{x} + b$. Trường hợp hệ số tương quan bằng 0, hai đại lượng là độc lập tuyến tính. Phương trình biểu diễn tương quan được tính như sau:

$$\rho_{\mathrm{xy}} = \frac{\mathrm{cov(x,y)}}{\sigma_x \sigma_y}$$

Trong hồi qui tuyến tính và logistic, hệ số tương quan thường được dùng để ranking mức độ quan trọng của biến trước khi thực hiện hồi qui. Trong các mô hình timeseries như ARIMA, GARCH chúng ta cũng xác định các tham số bậc tự do của phương trình hồi qui dựa trên hệ số tương quan giữa các chuỗi với độ trễ của nó.

1.2. Qui luật số lớn

Qui luật số lớn cho rằng khi một mẫu con có kích thước càng lớn được rút ra từ tổng thể thì trung bình của nó càng đại diện cho tổng thể. Phát biểu toán học của qui luật số lớn:

Xét n mẫu ngẫu nhiên $X_1, X_2, \ldots, X_n$ độc lập cùng tuân theo phân phối $\mathbf{N}(\mu, \sigma^2)$, với mọi số thực ϵ dương, xác suất để khoảng cách giữa trung bình tích lũy và kỳ vọng $P(\frac{X_1+X_2+\ldots+X_n}{n} - \mathrm{E(X)} > \epsilon) \to 0$ khi $n \to \infty$ hay biểu diễn lim:

$$\lim_{n\to\infty} P(|\frac{X_1+X_2+\ldots+X_n}{n} - \mathrm{E(X)}| \geq \epsilon) = 0$$

Để chứng minh định lý này ta cần sử dụng đến bất đẳng thức Markov: xác suất để một biến ngẫu nhiên X không âm lớn hơn a ($a > 0$) luôn nhỏ hơn kì vọng của biến ngẫu nhiên đó chia cho a. $P(\mathrm{X} \geq a) \leq \frac{\mathrm{E(X)}}{a}$

Chứng minh bất đẳng thức markov:

Do x không âm nên

$$\begin{aligned}
\mathrm{E(X)} &= \int_0^\infty xf(x)dx \\
&= \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\
&\geq \int_a^\infty xf(x)dx \\
&\geq \int_a^\infty af(x)dx \\
&= a\int_a^\infty f(x)dx \\
&= a.P(\mathrm{X} \geq a)
\end{aligned}$$

Từ đó suy ra $P(\mathrm{X} \geq a) \leq \frac{\mathrm{E(X)}}{a}$

Chứng minh qui luật số lớn:

$$P(|\frac{X_1+X_2+\ldots+X_n}{n} - \mathrm{E(X)}| \geq \epsilon) = P((\frac{X_1+X_2+\ldots+X_n}{n} - \mathrm{E(X)})^2 \geq \epsilon^2)$$

Đặt $\mathrm{Z} = (Y_n - \mathrm{E(X)})^2$. Áp dụng bất đẳng thức markov cho đại lượng không âm Z, ta có: $P(\mathrm{Z} \geq \epsilon^2) \leq \frac{\mathrm{E(Z)}}{\epsilon^2}$ Mặt khác khi n tiến tới ∞:

$$\mathrm{E}(Y_n) = \mathrm{E(X)}$$

Ở đây ta coi $X_1, X_2, \ldots, X_n$ là các biến độc lập. Khi đó:

$$\mathrm{Var}(Y_n) = \mathrm{Var}(\frac{X_1+X_2+\cdots+X_n}{n}) = \frac{n\mathrm{Var(X)}}{n^2} = \frac{\mathrm{Var(X)}}{n}$$

Do đó:

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{E(Z)} &= \lim_{n\to\infty} \mathrm{E}(Y_n - \mathrm{E(X)})^2 \\
&= \lim_{n\to\infty} \mathrm{E}(Y_n - \mathrm{E}(Y_n))^2 \\
&= \lim_{n\to\infty} \mathrm{Var}(Y_n) \\
&= \lim_{n\to\infty} \frac{\mathrm{Var(X)}}{n} = 0
\end{aligned}$$

Từ đó thế vào (1) ta suy ra:

$$\begin{aligned}
\lim_{n\to\infty} P(\mathrm{Z} \geq \epsilon^2) &\leq \lim_{n\to\infty} \frac{\mathrm{E(Z)}}{\epsilon^2} \\
&= \lim_{n\to\infty} \frac{\mathrm{Var(X)}}{n\epsilon^2} = 0
\end{aligned}$$

Mặt khác $P(\mathrm{Z} \geq \epsilon^2) \geq 0$ nên suy ra $\lim_{n\to\infty} P(\mathrm{Z} \geq \epsilon^2) = 0$. Suy ra điều phải chứng minh. Mấu chốt của chứng minh bất đẳng thức này là chúng ta phải phát hiện được tính chất $\mathrm{Var}(Y_n) = \frac{\mathrm{Var(X)}}{n}$ là một đại lượng tiến dần về 0 khi n tiến tới vô cùng.

1.3. Định lý giới hạn trung tâm

Đây là một định lý rất nổi tiếng và quan trọng trong xác suất thống kê. Định lý trung tâm (central limit theorem) cho rằng khi rút ra một mẫu con đủ lớn từ một mẫu tổng thể của biến X có trung bình và phương sai hữu hạn thì giá trị trung bình của mẫu con sẽ hội tụ về trung bình của mẫu tổng thể. Chính nhờ định lý này chúng ta có thể rút ra được các tính chất của một biến ngẫu nhiên nhờ thu thập một mẫu con với kích thước đủ lớn. Các thông tin có thể suy diễn ra đó là các tham số của phân phối thông kê (trung bình, phương sai) và ước lượng khoảng tin cậy. Thật vậy: Xét một biến ngẫu nhiên X tuân theo phân phối chuẩn $\mathbf{N}(\mu, \sigma^2)$. Lấy một mẫu con đủ lớn $S_X = X_1, X_2, \ldots, X_n$ có các quan sát độc lập với kích thước n từ tổng thể X. Khi đó giá trị trung bình của chuỗi $\bar{X}$ sẽ tuân theo qui luật phân phối chuẩn $\mathbf{N}(\mu, \frac{\sigma^2}{n})$.

Nếu đặt $\mathrm{Z} = \frac{\bar{\mathrm{X}}-\mu}{\sigma\sqrt{n}}$ ta sẽ thu được một biến Z tuân theo phân phối chuẩn hoá $\mathbf{N}(0, 1)$. Coi $\bar{\mathrm{X}}$ là trung bình và s_x^2 là phương sai của S_X.

Khoảng tin cậy $1 - \alpha$ của trung bình tổng thể có thể được ước lượng qua các trung bình và phương sai của S_X:

$$[\bar{\mathrm{X}} - u_{\alpha/2}\frac{s_x}{\sqrt{n}}, \bar{\mathrm{X}} + u_{\alpha/2}\frac{s_x}{\sqrt{n}}]$$

Phần 2 - xác suất

2.1. Khái niệm biến ngẫu nhiên:

Biến ngẫu nhiên là một đại lượng được sử dụng để đo lường kết quả của những sự kiện ngẫu nhiên. Chẳng hạn như $\mathbf{x} \in 1, 2, 3, 4, 5, 6$ là các biến cố trong phép đo kết quả các lần tung một xúc xắc 6 mặt đồng chất thì $\mathbf{x}$ được coi là biến ngẫu nhiên.

Trong xác suất thống kê có hai khái niệm biến ngẫu nhiên là biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục. Biến ngẫu nhiên rời rạc là đại lượng mà các giá trị của nó nằm trong một tập hợp cho trước. Trái lại, biến ngẫu nhiên liên tục có miền giá trị là tập con thuộc $\mathcal{R}$ (tập số thực), có thể hữu hạn hoặc không hữu hạn.

Hàm phân phối xác suất (*probability distribution function*) $p(x)$ của một biến ngẫu nhiên $\mathbf{x}$ rời rạc là một hàm số đo lường xác suất xảy ra sự kiện $p(\mathbf{x} = x)$ của một biến cố. Như vậy $1 \geq p(x) \geq 0$ và tổng xác suất của toàn bộ các khả năng trong không gian biến cố bằng 1, hay:

$$\sum_{x\in\mathcal{S}} p(x) = 1$$

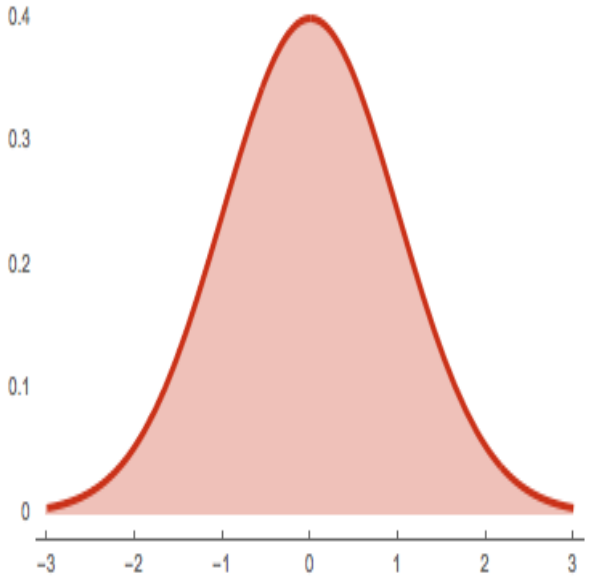
Trong đó $\mathcal{S}$ là không gian biến cố, chẳng hạn trường hợp tung đồng xu thì $\mathcal{S} = 1, 2, 3, 4, 5, 6$.

Khi biến ngẫu nhiên liên tục sẽ có vô số các giá trị có thể của $\mathbf{x}$. Vì vậy ta không thể biểu diễn khả năng xảy ra của toàn bộ sự kiện dưới dạng tổng xác suất rời rạc. Khi đó tích phân sẽ được sử dụng thay thế.

$$\int p(x)dx = 1$$

Trong trường hợp này thuật ngữ hàm mật độ xác suất (*probability density function* - pdf) để thể hiện $p(x)$ thay vì hàm phân phối xác suất (*probability distribution function*).

Như chúng ta đã biết tích phân của một hàm số $f(x)$ chính là diện tích nằm giữa đường cong đồ thị $y = f(x)$ và trục hoành. Như vậy, phần diện tích nằm dưới hàm mật độ xác suất $p(x)$ và trên trục hoành luôn có giá trị là 1. Chẳng hạn như đồ thị hàm mật độ xác suất của phân phối chuẩn như hình bên dưới:



> Hàm mật độ xác suất của phân phối chuẩn có phương trình $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ là đường cong có hình quả chuông đối xứng 2 bên. Giá trị hàm mật độ xác suất tại những điểm lùi về phía 2 đuôi trái và phải nhỏ dẫn và giá trị hàm mật độ xác suất tại vị trí trung tâm $x = \mu$ là lớn nhất. Phần diện tích màu hồng nằm dưới đường cong hàm mật độ xác suất và trục hoành có giá trị bằng 1.

2.2. Phân phối xác suất đồng thời.

Trường hợp trên là đối với không gian xác suất chỉ gồm 1 biến cố. Trên thực tế có nhiều biến cố xảy ra có mối liên hệ với nhau và đòi hỏi phát xét đến những không gian xác suất đồng thời của nhiều biến cố. Chúng ta sẽ thể hiện các xác suất đồng thời thông qua hàm phân phối xác suất đồng thời $p(x, y)$ biểu thị khả năng xảy ra đồng thời của cả 2 sự kiện x và y.

Tổng các khả năng xảy ra của các biến cố trong không gian các biến cố đồng thời luôn bằng 1. Điều đó có nghĩa là:

Nếu x, y rời rạc:

$$\sum_{x,y} p(x, y) = 1$$

Nếu x, y liên tục:

$$\int p(x, y) dx dy = 1$$

Nếu x rời rạc, y liên tục:

$$\sum_x \int p(x,y)dy = 1$$

2.3. Phân phối xác suất biên:

Nếu chúng ta cố định một biến cố và tính tổng (đối với biến rời rạc) hoặc tích phân (đối với biến liên tục) các xác suất chung $p(x,y)$ theo biến cố còn lại thì ta sẽ thu được hàm phân phối xác suất của theo một biến. Hàm phân phối xác suất này được gọi là xác suất biên (*marginal probability*) được tính như sau:

Biến rời rạc:

$$p(x) = \sum_y p(x,y)$$

$$p(y) = \sum_x p(x,y)$$

Biến liên tục:

$$p(x) = \int_y p(x,y)dy$$

$$p(y) = \int_x p(x,y)dx$$

2.4. xác suất có điều kiện:

xác suất của y theo điều kiện của x kí hiệu là $p(y|x)$ còn được gọi là xác suất hậu nghiệm (posterior probability) trong thống kê bayesian (bayesian statistic) có công thức như sau:

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

Xác suất hậu nghiệm cho ta biết khả năng xảy ra của một biến cố (biến cố y) trong điều kiện đã xét đến khả năng xảy ra của các biến cố khác (biến cố x).

Ngoài ra xác suất $p(x)$ còn được gọi là xác suất tiên nghiệm (prior probability), tức xác suất dựa trên niềm tin hoặc kinh nghiệm đã biết từ trước, trước khi các dấu hiệu xác suất (chính là điều kiện y) xuất hiện.

Từ công thức xác suất trên suy ra:

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$$

Ví dụ xác suất có điều kiện: xác suất chiến thắng (biến cố y) trong điều kiện tung được xúc sắc mặt 6 (biến cố x) là:

$$p(y|x=6) = \frac{p(x,y=6)}{p(x=6)}$$

Khi đó $p(x=6)$ thông thường sẽ bằng $\frac{1}{6}$ nếu khối xúc sắc là đồng chất chính là xác suất tiên nghiệm mà ta đã biết trước, ngay cả khi không cần đến điều kiện y là người đó đã chiến thắng.

Xác suất $p(y|x=6)$ là xác suất hậu nghiệm cho biết khả năng chiến thắng trong điều kiện đã biết tung được mặt $x=6$.

Trong các mô hình classification, xác suất dự báo đối với input là quan sát X sẽ là xác suất hậu nghiệm $P(Y=1|X)$ trong điều kiện mẫu có các đặc trưng mẫu là X.

2.5. Công thức bayes

Chúng ta có thể biểu diễn xác suất có điều kiện của biến cố y theo x dựa trên xác suất có điều kiện của biến cố x theo y.

$$\begin{aligned} p(y|x) &= \frac{p(x,y)}{p(x)} \\ &= \frac{p(x,y)}{\sum_y p(x,y)} \\ &= \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)} \end{aligned}$$

Ví dụ: Gọi y là biến cố khách hàng vỡ nợ, x là thu nhập khách hàng. Tính xác suất khách hàng vỡ nợ trong điều kiện khác hàng thu nhập dưới 10 triệu VND biết rằng $p(y=1)=0.01$ và xác suất khách hàng có thu nhập dưới 10 triệu trong điều kiện vỡ nợ và không vỡ nợ lần lượt là 0.9 và 0.05.

Lời giải: Từ điều kiện xác suất khách hàng có thu nhập dưới 10 triệu trong điều kiện vỡ nợ và không vỡ nợ lần lượt là 0.9 và 0.05 ta có $p(x<10|y=1)=0.9$ và $p(x<10|y=0)=0.05$.

Áp dụng công thức bayes:

$$\begin{aligned} p(y=1|x<10) &= \frac{p(x<10,y=1)}{p(x<10)} \\ &= \frac{p(x<10,y=1)}{\sum_y p(x<10,y)} \\ &= \frac{p(x<10|y=1)p(y=1)}{p(x<10|y=1)p(y=1)+p(x<10|y=0)p(y=0)} \\ &= \frac{0.9\times 0.01}{0.9\times 0.01+0.05\times 0.99} \\ &= \frac{0.009}{0.009+0.0495} = 0.153846 \end{aligned}$$

Điểm mấu chốt của công thức bayes đó là chúng ta có thể tính được xác suất có điều kiện khi đã biết phân phối xác suất ngược lại của điều kiện theo biến cố cần tính xác suất.

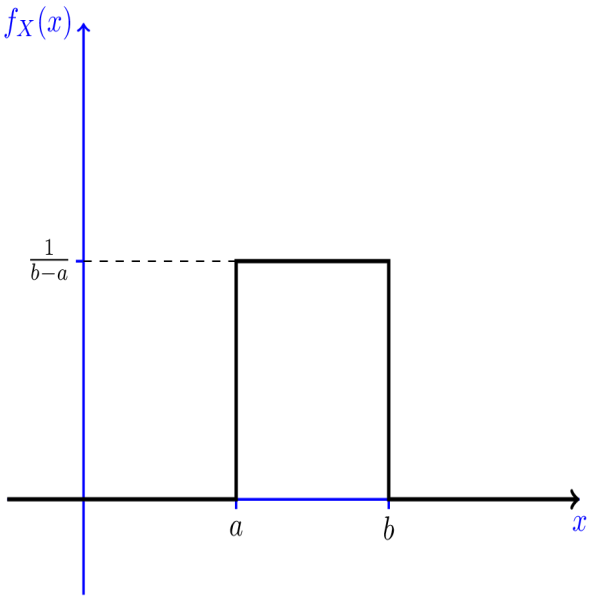
Phần 3 - Kiểm định và phân phối thống kê

3.1. Phân phối thống kê

Thống kê là bộ môn khoa học dựa trên các qui luật số lớn. Từ thời kì cổ đại các nhà toán học đã nhận ra một số qui luật của thống kê chẳng hạn như khi tung một đồng xu đồng chất thì xác suất nhận được các mặt xấp và ngửa đều bằng nhau và bằng 0.5. Chính qui luật đơn giản này đã hình

thành nên một phân phối nổi tiếng về xác suất là phân phối bernouli cho biết khả năng để xảy ra một biến cố K trong một phép thử ngẫu nhiên là $\lambda_K \in [0, 1]$. Từ định nghĩa về phân phối bernouli bạn đọc đã hiểu được phân phối là gì rồi chứ? Phân phối chính là đặc trưng cho sự phân bố của các biến ngẫu nhiên trên toàn tập xác định của nó. Trong tự nhiên có những qui luật phân phối tri phối hầu hết các nhân tố, chúng có thể mô hình hoá được và đúng với số lớn mà ta sẽ tìm hiểu ở bên dưới. Nhưng trước tiên chúng ta sẽ làm quen với các khái niệm trong phân phối.

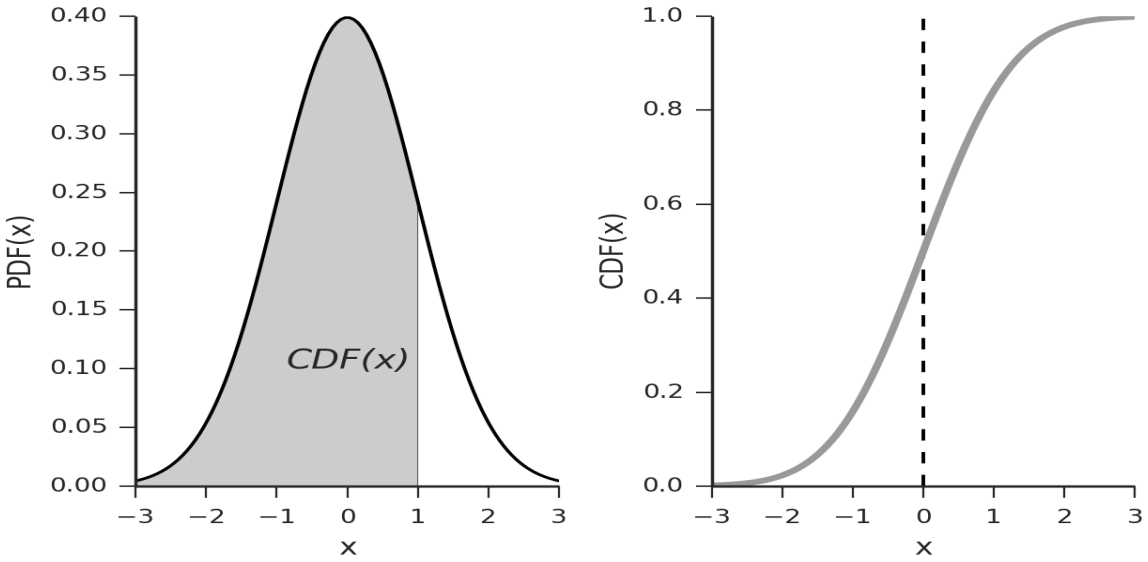
1. **Hàm mật độ xác suất và phân phối xác suất**:
   Trong thống kê chúng ta có rất nhiều các định dạng phân phối khác nhau. Trong đó có những phân phối cơ bản và thông dụng nhất bao gồm: phân phối chuẩn, t-student, Chi-square, Fisher, Bernouli, Phân phối nhị thức và Poission. Một phân phối được đặc trưng bởi 1 hàm số thể hiện giá trị của phân phối xảy ra tại mỗi một điểm. Tuỳ thuộc vào biến cố là liên tục hay rời rạc mà tên hàm phân phối của chúng có thể khác nhau. Trong trường hợp biến liên tục hàm đại diện cho một phân phối được gọi là `mật độ xác suất` (*pdf - probability density function*) và tên gọi hàm `phân phối xác suất` (*pmf - probability mass function*) được sử dụng với biến rời rạc.
   Giá trị hàm `phân phối xác suất` và `mật độ xác suất` được ký hiệu là $f_X(x)$ để biểu thị khả năng xảy ra của $\mathrm{X} = x$. Giá trị của hàm mật độ và hàm phân phối xác suất đều lớn hơn hoặc bằng 0 vì xác suất luôn không âm. Có một sự khác biệt giữa *pdf* và *pmf* đó là giá trị của *pdf* có thể lớn hơn 1 trong khi *pmf* luôn nhỏ hơn 1. Xảy ra điều này là bởi với các biến rời rạc thì tổng các xác suất trên toàn không gian biến cố bằng 1. Do đó giá trị của *pmf* không vượt qua 1. Trái lại, khi biến liên tục xác suất toàn miền được tính bằng tích phân của hàm *pdf*. Một hàm số lớn hơn 1 vẫn có thể cho giá trị tích phân nhỏ hơn 1 nên *pdf* hoàn toàn có thể có giá trị lớn hơn 1. Cụ thể hơn chúng ta có thể lấy ví dụ về phân phối đều (*uniform distribution*) là phân phối có miền xác định trên đoạn $[a, b]$ và xác suất của chúng luôn bằng nhau tại mọi điểm trên miền xác định.
   Phương trình hàm mật độ xác suất *pdf* có dạng: $f_X(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & x < a \text{ or } x > b \end{cases}$



Hình 2: Đồ thị hàm mật độ xác suất của phân phối đều trên đoạn $[a, b]$. Khi $b - a < 1$ thì hàm mật độ xác suất có thể lớn hơn 1.

1. **Hàm phân phối xác suất tích luỹ**:
   Hàm phân phối xác suất tích lũy (*cdf - cumulative distribution function*) tính toán xác suất xảy ra của một biến ngẫu nhiên trong một khoảng giá trị. Giá trị của hàm phân phối xác

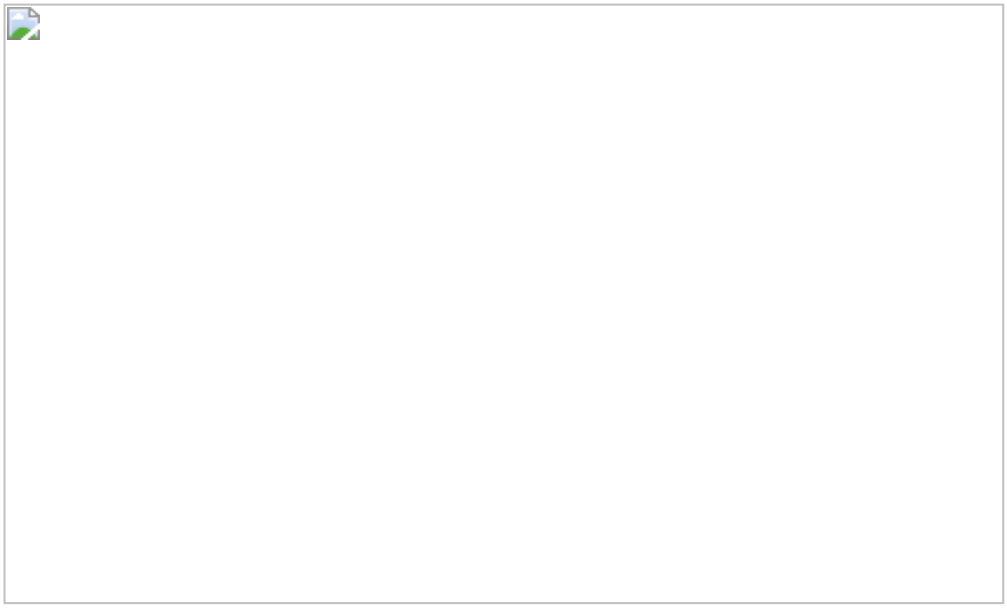
suất tích luỹ chính là tích phân của hàm mật độ xác suất trong trường hợp biến liên tục hoặc tổng của hàm phân phối xác suất trong trường hợp biến rời rạc. Do đó kí hiệu của nó thường là $F_X(x)$. Hàm *cdf* được biểu thị trên đồ thị như thế nào? Hẳn chúng ta còn nhớ khái niệm về tích phân đã từng học tại THPT, đây chính là phần diện tích nằm dưới đồ thị của hàm số và nằm trên trục hoành. Chẳng hạn trong phân phối chuẩn ta có $CDF(\mathrm{X} < 1)$ hoặc $F_X(1)$ chính là diện tích phần gạch chéo hình bên dưới. Khi đồ thị hóa hàm phân phối xác suất tích lũy ta thu được đồ thị như hình bên phải biểu diễn xác suất tích lũy theo X. Ta nhận thấy rằng $F_X(x)$ là một hàm đồng biến trên toàn miền giá trị.



Hình 3: Diện tích biểu diễn hàm phân phối xác suất $F_X(1)$ (phần gạch chéo).

3.2. Phân phối chuẩn.

Phân phối chuẩn là phân phối nổi tiếng nhất trong thống kê. Nó được tìm ra bởi nhà toán học Gaussian (hoàng tử của các nhà toán học) nên còn được gọi là phân phối Gaussian. Người ta từng ví rằng việc tìm ra qui luật phân phối chuẩn quan trọng giống như việc tìm ra 3 định luật của Newton trong vật lý cổ điển. Người Đức tự hào về phân phối chuẩn đến mức đã cho in hình quả chuông chuẩn trên tờ tiền của họ.



Hình 4: Hình ảnh phân phối chuẩn bên cạnh nhà toán học Gaussian trên đồng tiền Đức.

Quay trở lại lý thuyết, phân phối này được mô tả bởi hai tham số: trung bình μ và phương sai σ^2. Giá trị của μ là vị trí trung tâm của đáy phân phối có giá trị của hàm mật độ xác suất là cao nhất. Phân phối có độ rộng đáy càng lớn khi σ^2 lớn, điều này chứng tỏ khoảng giá trị của biến biến động mạnh, và ngược lại. Hàm mật độ xác suất của phân phối này được định nghĩa là:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Ứng dụng của phân phối chuẩn:

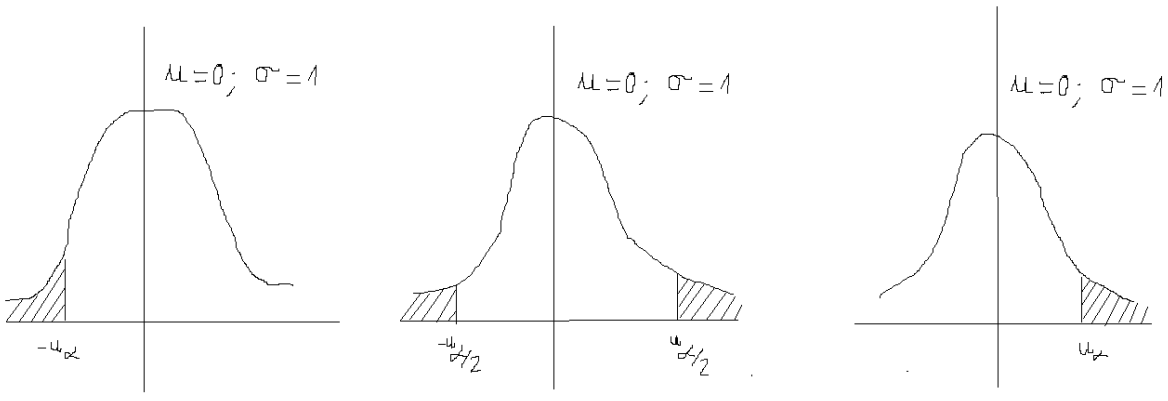
Phân phối chuẩn thường được sử dụng nhiều trong thống kê để ước lượng khoảng tin cậy, tính toán xác suất và kiểm định giả thuyết thống kê. Chúng ta sẽ lần lượt đi qua các ứng dụng này.

3.2.1. Ước lượng khoảng tin cậy.

Khoảng tin cậy của một biến ngẫu nhiên là một miền giá trị mà ta tin rằng khả năng để biến ngẫu nhiên rơi vào miền đó với mức độ tin cậy là bao nhiêu. Mức độ tin cậy càng cao thì khẳng định của chúng ta càng chắc chắn.

Ví dụ về khoảng tin cậy: Một người nông dân trồng vải thiều muốn ước tính giá vải thiều vào giữa vụ với mức độ tin cậy là 95% sẽ rơi vào khoảng nào. Biết rằng giá vải thiều năm trước có trung bình là 30 nghìn đồng/kg và phương sai là 5 nghìn đồng/kg.

Như vậy một khoảng tin cậy luôn gắn liền với một mức độ tin cậy. Trong thống kê, nếu một biến ngẫu nhiên $\mathrm{X} \sim \mathbf{N}(\mu, \sigma^2)$. Với mức độ tin cậy $(1-\alpha)$, thông thường $\alpha = 0.05$, ta sẽ có các định nghĩa về khoảng tin cậy sau đây:



(a) Khoảng tin cậy trái (b) Khoảng tin cậy 2 phía (c) Khoảng tin cậy phải

Hình minh họa các `khoảng tin cậy trái`, `khoảng tin cậy 2 phía` và `khoảng tin cậy phải` tại mức ý nghĩa $1-\alpha$ và biến ngẫu nhiên phân phối $\mathrm{X} \sim \mathbf{N}(\mu, \sigma^2)$.

- Khoảng tin cậy 2 phía: Là khoảng tin cậy đối xứng qua trung bình. Cho ta biết khoảng biến động quanh giá trị trung bình của một biến ngẫu nhiên tương ứng với mức độ tin cậy $(1-\alpha)$. Ví dụ: Giá vải thiều giao động trong khoảng nào với mức độ tin cậy $(1-\alpha)$?

$$\mu - u_{\alpha/2}.\sigma \leq \mathrm{X} \leq \mu + u_{\alpha/2}.\sigma$$

- Khoảng tin cậy trái: Cho ta biết điểm cận dưới mà biến ngẫu nhiên có khả năng lớn hơn với mức độ tin cậy tương ứng. Ví dụ: Giá vải thiều thấp nhất là bao nhiêu với mức độ tin cậy $(1-\alpha)$?

$$X \geq \mu - u_\alpha . \sigma$$

- Khoảng tin cậy phải: Ngược lại với tin cậy bên trái, cho ta biết điểm cận trên mà biến ngẫu nhiên có khả năng nhỏ hơn với mức tin cậy tương ứng. Ví dụ: Giá vải thiều cao nhất là bao nhiêu với mức độ tin cậy $(1-\alpha)$?

$$X \geq \mu + u_\alpha . \sigma$$

Bạn đọc đã hình dung ra được ứng dụng của khoảng tin cậy rồi chứ? Các khoảng tin cậy được tính dựa trên các tham số phân phối trung bình và phương sai và giá trị tới hạn.

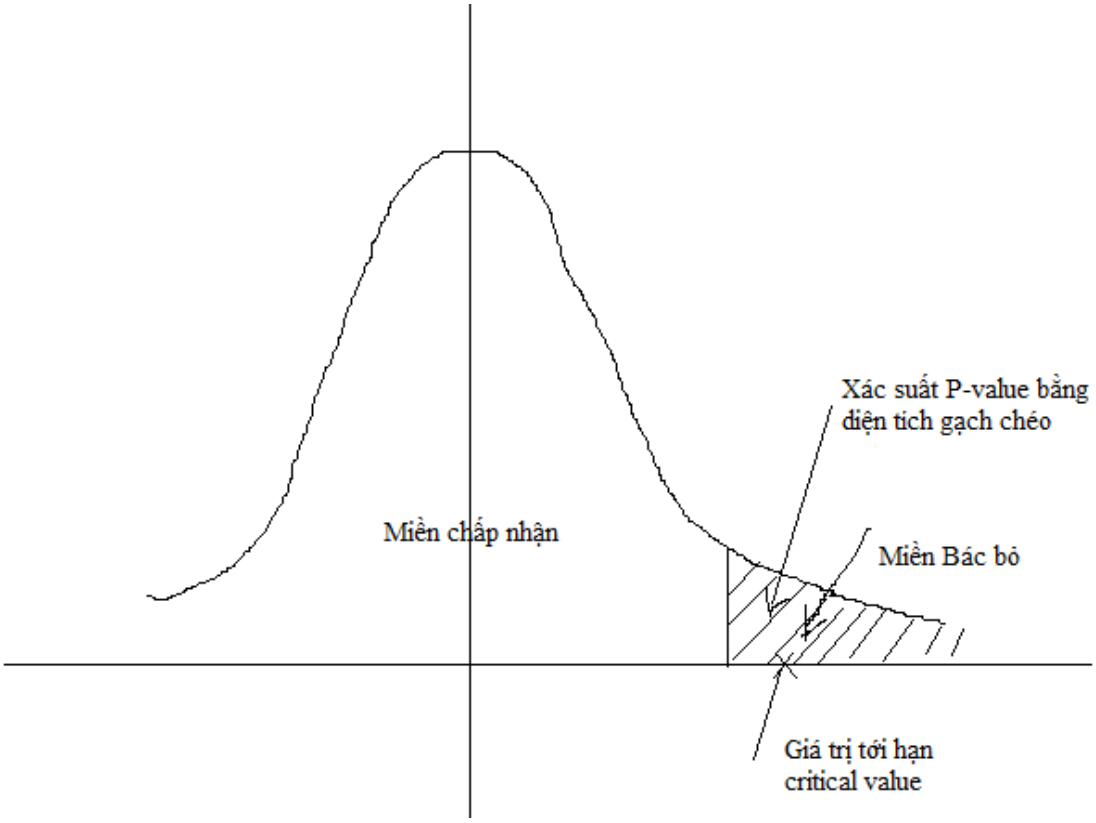
Vậy giá trị tới hạn là gì?

Trong công thức trên, u_α được gọi là giá trị tới hạn (critical value) của phân phối. Có thể coi u_α là hàm ngược của hàm phân phối xác suất tích luỹ. Tức là nếu ta có: $P(\mathrm{X} \geq a) = \alpha$ thì giá trị $u_\alpha = a$.

Giải thích dễ hiểu hơn thì u_α chính là ngưỡng thấp nhất để xác suất biến ngẫu nhiên có giá trị lớn hơn ngưỡng đó là α.

3.2.2. Kiểm định giả thuyết thống kê

Giả thuyết thống kê được xây dựng để kiểm chứng các giả thuyết được đặt ra đối với một biến ngẫu nhiên. Một giả thuyết thống kê sẽ bao gồm 2 vế đối nghịch là giả thuyết khống (null hypothesis) kí hiệu là H_0 và giả thuyết thay thế (alternative hypothesis) kí hiệu là H_1. Giả thuyết khống là giả thuyết không có ở hiện tại và ta đang tìm cách kiểm chứng. Giải thuyết thay thế là giả thuyết đối nghịch lại đối với H_0. Trong kiểm định giả thuyết thống kê chúng ta chấp nhận một ngưỡng mắc sai lầm cho phép tối đa đối với giả thuyết H_0 là xác suất P-value (Probability value) (thông thường là 0.05). Phân phối xác suất của X được chia làm 2 miền đối nghịch là miền chấp nhận và miền bác bỏ đối với giả thuyết H_0. Miền bác bỏ có xác suất xảy ra là `P-value`. Đây cũng chính là xác suất tối đa được phép mắc sai lầm của H_0.



Các kiểm định thống kê sẽ tính ra giá trị ngưỡng hoặc giá trị tới hạn `critical -value` và xét xem giá trị này rơi vào miền chấp nhận hay bác bỏ để đưa ra quyết định đối với giả thuyết H_0. Như trong hình minh họa trên thì giá trị tới hạn rơi vào miền bác bỏ nên ta sẽ bác bỏ giả thuyết H_0 với mức ý nghĩa là P-value.

Trong kiểm định thống kê có 3 cặp giải thuyết thống kê đó là: `Giả thuyết dấu bằng, giả thuyết lớn hơn, giả thuyết nhỏ hơn` . Chúng ta sẽ lần lượt đi qua 3 giả thuyết này.

1. Giả thuyết dấu bằng:

Giả sử có một giả thuyết cho rằng trung bình của x là μ_0. Kiểm chứng giả thuyết này với mức độ tin cậy là 95% (tương ứng với P-value = 5%). Khi đó ta gọi đây là trường hợp kiểm định giả thuyết dấu bằng. Kiểm định này gồm 2 vế giả thuyết như sau:

$$\begin{cases} H_0 : \bar{\mathrm{X}} = \mu \\ H_1 : \bar{\mathrm{X}} \neq \mu \end{cases}$$

Bác bỏ giả thuyết dựa trên miền chấp nhận:

Với mức ý nghĩa $(1-\alpha)$ thì miền chấp nhận giả thuyết H_0 là:

$$\mathcal{D} = \{\mathrm{X} \sim \mathbf{N}(\mu, \sigma^2) | \mu - u_{\alpha/2}.\sigma \leq \mathrm{X} \leq \mu + u_{\alpha/2}.\sigma\}$$

Trên thực tế xác xuất tính theo phân phối chuẩn

$$P(\mu - u_{\alpha/2}.\sigma \leq \mathrm{X} \leq \mu + u_{\alpha/2}.\sigma) = 1 - \alpha$$

Mức ý nghĩa ở đây có thể hiểu là khả năng chắc chắn để giả thuyết H_0 xảy ra. Nếu mức ý nghĩa là 95%, ta có thể khẳng định chắc rằng khả năng rơi vào miền $\mathcal{D}$ của X là 95%.

Bác bỏ giả thuyết dựa trên xác suất:

Một cách khác để kết luận giả thuyết dấu bằng là dựa trên xác suất `P-value` được tính ra để X **không rơi** vào miền $\mathcal{D}$. Xác suất này nếu nhỏ hơn hoặc bằng α thì ta sẽ bác bỏ H_0.

$$1 - P(\bar{\mathrm{X}} - u_{\alpha/2}.\sigma \leq \mathrm{X} \leq \bar{\mathrm{X}} + u_{\alpha/2}.\sigma) = \mathrm{P\text{-}value}$$

Nếu $\mathrm{P\text{-}value} <= \alpha$ thì bác bỏ H_0. Chúng ta thường thấy kiểm định theo xác suất trong các kết quả từ mô hình hồi qui thống kê. Cột `P-value` thường được so sánh với 0.05 để kết luận chấp nhận hay bác bỏ giả thuyết.

Lưu ý khi xác định chấp nhận hay bác bỏ một giả thuyết ta luôn phải đi kèm với điều kiện kết luận ở mức ý nghĩa bao nhiêu %.

2. Giả thuyết lớn hơn:

Cặp giả thuyết lớn hơn dùng để kiểm chứng một nhận định giá trị của một biến lớn hơn một hằng số nào đó. giả thuyết này sẽ khác với cặp giả thuyết dấu bằng ở chỗ trong phát biểu của nó giả thuyết H_1 là một dấu lớn nhỏ hơn thay vì dấu khác. Biểu diễn cặp giả thuyết lớn hơn như sau:

$$\begin{cases} H_0 : \mathrm{X} = \mu \\ H_1 : \mathrm{X} < \mu \end{cases}$$

Với mức ý nghĩa $(1 - \alpha)$, miền chấp nhận giả thuyết H_0 là

$$\mathcal{D} = \{\mathrm{X} \sim \mathbf{N}(\mu, \sigma^2) | X \geq \mu - u_\alpha.\sigma\}$$

xác suất

$$P(X \geq \mu - u_\alpha.\sigma) = 1 - \alpha$$

3. Giả thuyết nhỏ hơn:

Được sử dụng để kiểm chứng một nhận định giá trị của một biến nhỏ hơn một hằng số nào đó. Hoàn toàn tương tự như cặp giả thuyết lớn hơn ta có biểu diễn của cặp giả thuyết nhỏ hơn.

$$\begin{cases} H_0 : \mathrm{X} = \mu \\ H_1 : \mathrm{X} > \mu \end{cases}$$

Với mức ý nghĩa $(1 - \alpha)$, miền chấp nhận giả thuyết H_0 của giả thuyết này là

$$\mathcal{D} = \{\mathrm{X} \sim \mathbf{N}(\mu, \sigma^2) | X \leq \mu - u_\alpha.\sigma\}$$

xác suất

$$P(X \leq \mu - u_\alpha.\sigma) = 1 - \alpha$$

3.3. t-student.

Như chúng ta đã biết hầu hết các mẫu ngẫu nhiên với kích thước đủ lớn đều tuân theo qui luật phân phối chuẩn. Tuy nhiên nhược điểm của phân phối chuẩn đó là chúng ta phải biết trước được các tham số về kì vọng và phương sai của tổng thể thì mới xác định được hình dạng của phân phối. Trong khi không phải khi nào cũng đo lường được những tham số này vì tổng thể là quá lớn. Do đó một phân phối khác được phát triển có hình dạng gần tương tự như phân phối chuẩn hoá nhưng ứng dụng trên phương sai và độ lệch chuẩn của mẫu thay vì tổng thể, đó chính là phân phối t-student.

t-student là phân phối thuộc họ các phân phối liên tục được phát triển trong quá trình ước lượng trung bình của các chuỗi phân phối chuẩn nhưng kích thước mẫu nhỏ và phương sai của tổng thể là chưa xác định. t-student được sử dụng chủ yếu trong các trường hợp sau:

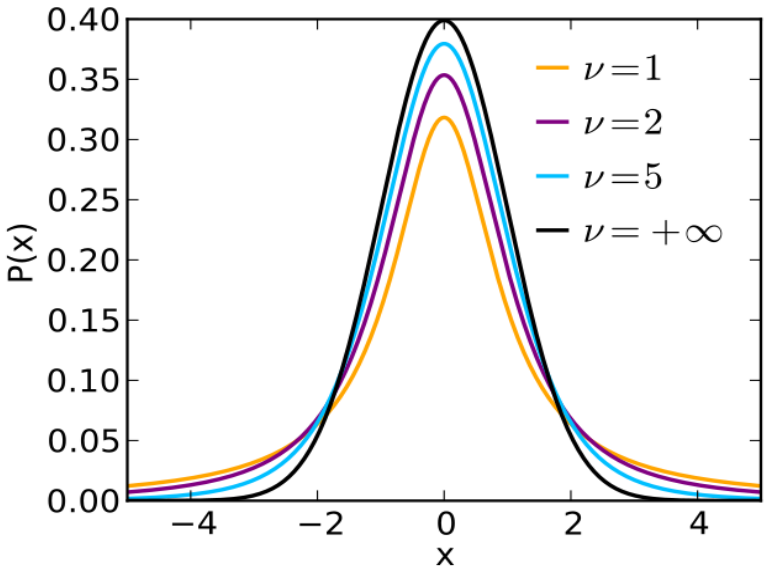
- Kiểm định về khác biệt giữa 2 trung bình mẫu.

- Tìm khoảng tin cậy về sự khác biệt giữa 2 trung bình mẫu.
- Sử dụng tìm khoảng tin cậy của các hệ số ước lượng và tính P-value của các hệ số ước lượng trong hồi qui tuyến tính.

Phát biểu của phân phối t-student như sau:

Nếu ta lấy một mẫu con $X_1, X_2, \ldots, X_n$ con kích thước n từ tổng thể của biến ngẫu nhiên X phân phối chuẩn có kì vọng μ nhưng chưa biết về phương sai của nó. Khi đó sai số của các phần tử trong mẫu so với μ sau khi nhân với $\frac{1}{S\sqrt{n-1}}$ là đại lượng tuân theo qui luật phân phối t-student với bậc tự do $n-1$. Trong đó S là phương sai của mẫu. Tức là: $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_i)^2}{n-1}$.

Trong trường hợp đã biết phương sai σ^2 của tổng thể: $\mathrm{Z} = \frac{\mathrm{X}-\mu}{\sigma\sqrt{n}} \sim \mathbf{N}(0,1)$ Trong trường hợp chưa biết phương sai tổng thể: $\mathrm{Z} = \frac{\mathrm{X}-\mu}{S\sqrt{n-1}} \sim \mathbf{T}(n-1)$ Do là một phân phối thay thế cho phân phối chuẩn hoá trong trường hợp chưa xác định phương sai tổng thể nên hình dạng của phân phối t-student cũng gần như phân phối chuẩn phối chuẩn hoá. Trên thực tế nếu $\mathrm{Z} \sim \mathbf{T}(n-1)$ thì $\mathrm{E}(\mathrm{Z}) = 0$ và $\mathrm{Var}(\mathrm{Z}) = \frac{n}{n-2}$ (với $n > 2$, còn lại không xác định). Trong trường hợp bậc tự do vô cùng lớn, phân phối t-student hội tụ về phân phối chuẩn hoá.

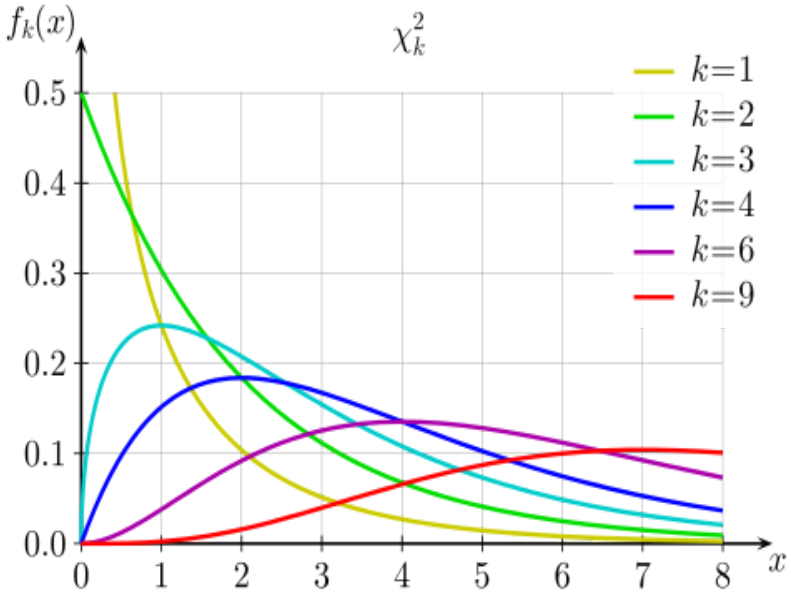


Hình 5: Hình dạng của phân phối t-student với bậc tự do lần lượt là $1, 2, 5, +\infty$

3.4. Phân phối Chi-square

Phân phối chi-square kí hiệu là χ^2 là một phân phối có bậc tự do. Một phân phối chi-square được tạo thành từ tổng bình phương của các phân phối chuẩn hóa mà bậc tự do của nó chính bằng số lượng các phần tử trong tổng. Hay nói cách khác: $X_1, X_2, \ldots, X_n$ là tợp hợp gồm n biến ngẫu nhiên, độc lập tuyến tính và phân phối chuẩn hóa thì biến $\mathrm{Y} = \sum_{i=1}^{n} X_i^2$ là một phân phối chi-square với bậc tự do n. Ta kí hiệu $\mathrm{Y} \sim \chi_n^2$.

Bên dưới là đồ thị của hàm mật độ xác suất của phân phối chi-square.



Hình 6: Hàm mật độ xác suất (*pdf*) của phân phối chi-square với bậc tự do từ 1 đến 9

Ta nhận thấy chi-square là một phân phối lệch trái. Khi bậc tự do của nó càng nhỏ thì đồ thị lệch trái đồng thời phần đuôi phía bên phải càng mỏng và trái lại.

Phân phối chi-square là một trong những phân phối phổ biến nhất trong suy diễn thống kê, thường được sử dụng trong các kiểm định giả thuyết và tìm khoảng tin cậy. Một số ứng dụng cụ thể của chi-square:

- Kiểm tra tính phù hợp (*goodness of fit*) của một phân phối thực nghiệm theo một phân phối lý thuyết. Chẳng hạn chúng ta có một chuỗi thực nghiệm O và một chuỗi lý thuyết phân phối chuẩn E có cùng kích thước mẫu n. Chúng ta nghi ngờ rằng O và E có cùng phân phối. Khi đó tổng bình phương sai số của E và O sẽ tuân theo phân phối chi-square bậc tự do n. $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$
  - χ^2: Gía trị kiểm định của giả thuyết 2 chuỗi có cùng phân phối, tuân theo phân phối chi-square bậc tự do $n-1$ (Lưu ý đối với mẫu con thì bậc tự do là $n-1$ còn đối với tổng thể bậc tự do là n).
  - O_i: quan sát thứ i của chuỗi thực nghiệm O.
  - E_i: quan sát thứ i của chuỗi lý thuyết nghiệm E.
  - n: Số lượng các quan sát.

  Để kết luận 2 chuỗi có cùng phân phối hay không, ta sẽ so sánh gía trị kiểm định của giả thuyết χ^2 với giá trị tới hạn $\chi^{2(1-\alpha)}_{n-1}$.
  Ví dụ: Một súc sắc 6 mặt được ném 60 lần. Số lần xuất hiện các mặt 1, 2, 3, 4, 5 và 6 lần lượt là 5, 8, 9, 8, 10 và 20. Kiểm định giả thuyết rằng có sự khác biệt về khả năng nhận được các mặt theo kểm định Pearson chi-squared ở mức ý nghĩa 95%?
  Kì vọng là khả năng nhận được các các mặt của xác suất là như nhau, do đó mỗi mặt được dự kiến xuất hiện là bằng nhau và bằng 60/6 = 10. Các kết quả được lập bảng như sau:

i	O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	5	10	−5	25	2.5
2	8	10	−2	4	0.4
3	9	10	−1	1	0.1
4	8	10	−2	4	0.4

5	10	10	0	0	0
6	20	10	10	100	10
				Sum	13.4

Bậc của tự do chính là $n - 1 = 5$. Gía trị tới hạn đuôi lớn hơn của phân phối chi-square tại mức tin cậy 95% được cho ở bảng bên dưới

Degrees of freedom	**Probability less than the critical value**				
	0.90	***0.95***	**0.975**	***0.99***	**0.999**
5	9.236	*11.070*	12.833	*15.086*	20.515

Gía trị kiểm định giả thuyết là 13.4 vượt qua giá trị tới hạn tại mức ý nghĩa 95%. Do đó bác bỏ giả thuyết dấu bằng H_0 và kết luận rằng xác suất khả năng xảy ra các mặt của súc sắc là khác nhau tại mức ý nghĩa 95%.

- Kiểm tra tính phụ thuộc giữa các biến phân loại dựa trên bảng cross table.
  Ví dụ: Để dễ minh họa tôi xin lấy ví dụng từ wikipedia. Chúng ta có số liệu về 100 học sinh được lựa chọn ngẫu nhiên theo 2 tiêu chí giới tính (Sex) và tay thuận (Handedness) được rút ra từ một tổng thể rất lớn một cách ngẫu nhiên. Bảng thông kê cross table cho thấy như dưới:

Handed-ness / **Sex**	**Right handed**	**Left handed**	**Total**
Male	43	9	52
Female	44	4	48
Total	87	13	100

Để kiểm tra xem liệu giới tính có ảnh hưởng lên tay thuận của học sinh hay không chúng ta có thể sử dụng kiểm định Pearson chi-squared tìm ra sự khác biệt giữa các nhóm tay thuận tay trái và tay phải theo giới tính.

- Ước lượng khoảng tin cậy cho phương sai của một chuỗi các độ lệch chuẩn.
  Thường được áp dụng trong tìm khoảng tin cậy trong phân tích chuỗi thời gian. Thu thập số liệu biến thiên tỷ suất lợi nhuận của các mã chứng khoán trong vòng 36 tháng ta sẽ thu được một chuỗi các độ lệch chuẩn $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{36}^2$. Tìm khoảng tin cậy 95% độ giao động của chuỗi chứng khoán trong tháng tiếp theo.

3.5. Phân phối Fisher

Phân phối Fisher rất thường xuyên xuất hiện trong kinh tế lượng vì ứng dụng trong việc tìm sự khác biệt giữa 2 phương trình hồi qui, và phân tích phương sai ANOVA. Bởi vì được nghĩ ra bởi nhà thống kê học nổi tiếng Fisher người được coi là đặt nền móng cho ngành thống kê hiện đại nên tên của phân phối được đặt theo tên ông. Phân phối Fisher được xây dựng dựa trên một phép chia của 2 đại lượng ngẫu nhiên tuân theo qui luật phân phối có bậc tự do. Do đó một phân phối Fisher đặc trưng bởi 2 bậc tự do, một của tử số và một của mẫu số.

Một số tính chất của phân phối Fisher:

- Nếu $\mathrm{X} \sim \mathbf{T}(n)$ thì $\mathrm{X}^2 \sim \mathbf{F}(1, n)$. Đây chính là lý do tại sao kiểm định giá trị của một hệ số ước lượng trong phương trình hồi qui có ý nghĩa thống kê hay không vừa có thể được thực hiện qua phân phối Fisher và phân phối t-student mà kết quả thu được là tương đương.
- Phân phối Fisher và phân phối chi-square có mối quan hệ gần gũi. Nếu 2 biến ngẫu nhiên $\text{X} \sim \chi^{2}{d_1}và\text{Y} \sim \chi^{2}{d2}$ thì khi đó thương giữa chúng là $\frac{\mathrm{X}/d_1}{\mathrm{Y}/d_2} \sim \mathbf{F}(d_1, d_2)$

- Nếu $\mathrm{X} \sim \mathbf{F}(d_1, d_2)$ thì $\mathrm{X}^{-1} \sim \mathbf{F}(d_2, d_1)$

Ví dụ về ứng dụng của kiểm định Fisher về ý nghĩa của các hệ số ước lượng trong phương trình hồi qui như sau:

Một tập hợp gồm p biến độc lập $\mathrm{X}_1, \ldots, \mathrm{X}_p$ và biến phụ thuộc Y. Hồi qui toàn bộ p biến giải thích ta thu được phương trình hồi qui:

$$a_0 + a_1\mathrm{X}_1 + \cdots + a_p\mathrm{X}_n + \epsilon = \mathrm{Y}$$

với a_i là các hệ số tự do, ϵ là thành phần đại diện cho sai số ngẫu nhiên.

Phương trình hồi qui trên có tổng bình phương sai số (RSS - residual sum square) là

$$\mathrm{RSS}_1 = \sum_{i=1}^{n} \epsilon_i^2$$

với n là số quan sát.

Kiểm tra hệ số p-value của ước lượng cho thấy các biến từ $X_q, X_{q+1}, \ldots, X_p, q < p$ không có ý nghĩa thống kê (p-value > 0.05). Loại các biến này ra khỏi phương trình hồi qui ta thu được phương trình hồi qui thứ 2.

$$a_0 + a_1\mathrm{X}_1 + \cdots + a_q\mathrm{X}_q + u = \mathrm{Y}$$

Phương trình này có tổng bình phương sai số là

$$\mathrm{RSS}_2 = \sum_{i=1}^{n} u_i^2$$

Kiểm định giả thuyết rằng các hệ số $a_q, a_{q+1}, \ldots, a_p$ không có ý nghĩa thống kê.

Khi đó chúng ta có cặp giải thuyết kiểm định:

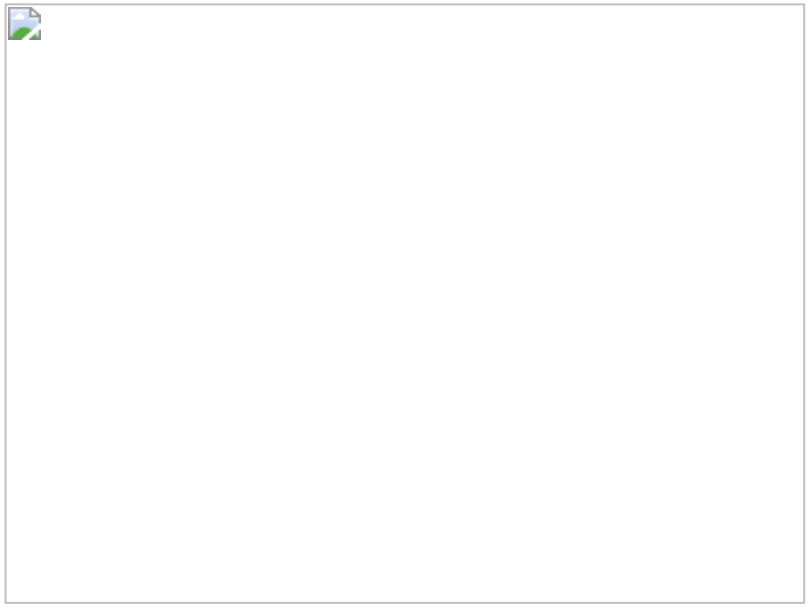
$$\begin{cases} H_0 : a_q = a_{q+1} = \cdots = a_p = 0 \\ H_1 : \sum_{i=q}^{p} a_i^2 > 0 \end{cases}$$

Việc chấp nhận giả thuyết H_0 tương đương với việc chấp nhận rằng 2 phương trình hồi qui như nhau. Do đó ta qui bài toán về kiểm định $\mathrm{RSS}_1 = \mathrm{RSS}_2$. Ta nhận thấy RSS_1 và RSS_2 đều là những phân phối chi-square nên thương của chúng sẽ có dạng một phân phối fisher. Ý tưởng là chúng ta cần tạo ra một phân phối fisher có thể dùng để tính toán giá trị tới hạn và đối chiếu với giá trị kiểm định. Đó chính là:

$$\mathrm{F} = \frac{(\mathrm{RSS}_1 - \mathrm{RSS}_2)/(p-q)}{\mathrm{RSS}_2/(n-p)}$$

có phân phối $\mathbf{F}(p-q, n-p)$.

Để bác bỏ H_0 với mức tin cậy 95% ta cần $\mathrm{F} > F_{0.05}(p-q, n-p)$ và trái lại. Chúng ta có thể biểu diễn miền bác bỏ giả thuyết H_0 là phần diện tích ở phía đuôi bên phải như hình bên dưới:



Hình 7: Miền bác bỏ giả thuyết H_0 với mức độ tin cậy 95% của kiểm định fisher.