Kiểm định giả thuyết thống kê

1. Kiểm định cổ điển:

- Giả sử ta xét bài toán về một tham số θ mà ta chưa biết. Ta chỉ biết nó thuộc về tập Θ.
- Giả sử tập tham số Θ có thể được phân hoạch thành hai tập tham số nhỏ hơn:

$$\Theta = \Theta_1 \cup \Theta_2$$

- Hãy xét khả năng tham số theta sẽ thuộc vào tập nào.

- Giả sử ta có hai giả thuyết:

  $H_0: \theta \in \Theta_1$ (Đây là giả thuyết 0)

  $\mathrm{H}_1: \theta \in \Theta_2$ (Đây là giả thuyết nâng cao)

- Câu hỏi đặt ra: làm thế nào để biết giả thuyết nào chính xác.

- Nếu ta bác bỏ giả thuyết H_0 nhưng giả thuyết này là đúng, thì lúc này ta sẽ gọi bước bác bỏ sai này là *lỗi loại 1.*
- Nếu ta chấp nhận giả thuyết H_0 nhưng giả thuyết này là giả thuyết sai, thì ta sẽ gọi bước bác bỏ sai này là *lỗi loại 2.*
- Các bước bác bỏ sẽ dựa trên biến quan sát:

$$\mathrm{R} = \{\mathrm{x};\ \text{biết x khiến cho giả thuyết } \mathrm{H}_0 \text{ bị bác bỏ}\}$$

- Phương pháp truyền thống (tần suất) xác định xác suất lỗi loại 1:

$$\text{Tính } \mathrm{P}(\mathrm{R}|\theta)$$

- Với lỗi loại 2, ta sẽ:

$$\text{Tính } 1 - \mathrm{P}(\mathrm{R}|\theta)$$

- Ta thấy rằng nếu xác suất lỗi loại một càng thấp thì khả năng lỗi loại 2 càng cao.
- Ta thường chọn R trước, sao cho xác suất gặp lỗi loại 2 càng nhỏ càng tốt, biết rằng xác suất lỗi loại 1 luôn được giới hạn ở một mức α cố định nào đó.
- α còn được gọi là test size.
- Phương pháp cổ điển không được trực tiếp và nhanh chóng vì phải xác định bộ R, như một bước trung gian.
- Vì vậy ta cần một phương pháp khác trực tiếp hơn.

2. Phương pháp kiểm định Bayes:

- Với trường phái Bayes, ta sẽ sử dụng những tính toán trực tiếp như sau:

$$\text{Tính } \mathrm{p}_0 = \mathrm{p}(\theta \in \Theta_0|\mathrm{D}) \text{ và } \mathrm{p}_1 = \mathrm{p}(\theta \in \Theta_1|\mathrm{D})$$

- Mục đích của chúng ta sẽ là so sánh trực tiếp H_0 và H_1.
- Chú ý rằng: $\mathrm{p}_0 + \mathrm{p}_1 = 1$ do không gian mẫu Θ bằng với $\Theta_0 \cup \Theta_1$.

- Khi áp dụng trường phái Bayes ta để ý rằng cần phải có xác suất tiên nghiệm.
- Trong trường hợp này xác suất tiên nghiệm ta cần xem xét là:

$$\pi_0 = \mathrm{p}(\theta \in \Theta_0) \text{ và } \pi_1 = \mathrm{p}(\theta \in \Theta_1)$$

- Lưu ý rằng $\pi_0 + \pi_1 = 1$.

- Từ đó, ta cũng có các đại lượng mới được gọi là *độ lệch tiên nghiệm của* H_0 *so với* H_1 và *độ lệch hậu nghiệm của* H_0 *so với* H_1.
- Độ lệch tiên nghiệm:

$$\frac{\pi_0}{\pi_1}$$

- Độ lệch hậu nghiệm:

$$\frac{\mathrm{p}_0}{\mathrm{p}_1}$$

- Ta cũng thấy rằng nếu độ lệch tiên nghiệm càng tiến về 1, ta sẽ coi H_0 và H_1 gần như nhau.
- Nếu chênh lệch tiên nghiệm lớn thì ta sẽ coi trọng giả thuyết H_0 hơn giả thuyết H_1.
- Nếu chênh lệch tiên nghiệm càng nhỏ thì sẽ ít coi trọng giả thuyết H_0 hơn giả thuyết H_1.
- Ta cũng đó điều tương tự với độ lệch hậu nghiệm.

Chú ý:

- Lưu ý rằng trong cả hai hướng tiếp cận xác suất truyền thống lẫn Bayes, ta phải lưu ý rằng H_0 luôn là đối tượng được ưu tiên xem xét trước.
- Bài toán kiểm định giả thuyết thống kê luôn phải xoay quanh giả thuyết H_0.

- Sau khi có hai đại lượng độ lệch tiên nghiệm và hậu nghiệm, ta sẽ đến với một đại lượng mới, được gọi là đại lượng *phân rã Bayes* với ưu tiên dành cho H_0 so với H_1*:*

$$\mathrm{B} = \frac{\frac{\mathrm{p}_0}{\mathrm{p}_1}}{\frac{\pi_0}{\pi_1}} = \frac{\mathrm{p}_0 . \pi_1}{\mathrm{p}_1 . \pi_0}$$

- Đại lượng B hoàn toàn có thể được hiểu theo nghĩa là “độ chênh lệch ưu tiên của giả thuyết H_0 so với giả thuyết H_1 được *cho bởi bộ dữ liệu*”.
- Lưu ý rằng ta hoàn toàn có thể tìm ra công thức tính xác suất hậu nghiệm của p_0 dựa trên xác suất tiên nghiệm π_0 và đại lượng chênh lệch Bayes B.

- Công thức có thể viết lại như sau:

$$\frac{p_0}{p_1} = B.\frac{\pi_0}{\pi_1}$$

$$\Rightarrow\ p_0 = \frac{1}{1 + \frac{\pi_0}{\pi_1} B^{-1}} = \frac{1}{1 + \frac{\pi_0}{1 - \pi_0} B^{-1}}$$

- Trong trường hợp $\Theta_0 = \{\theta_0\}, \Theta_1 = \{\theta_1\}$, ta sẽ có:

$$\mathrm{p}_0 \sim \pi_0 \mathrm{p}(\mathrm{x}|\theta_0) \text{ và } \mathrm{p}_1 \sim \pi_1 \mathrm{p}(\mathrm{x}|\theta_1)$$

- Và như vậy ta được:

$$\frac{\mathrm{p}_0}{\mathrm{p}_1} = \frac{\pi_0 \mathrm{p}(\mathrm{x}|\theta_0)}{\pi_1 \mathrm{p}(\mathrm{x}|\theta_1)}$$

- Tuy nhiên trong trường hợp Θ_0 hoặc Θ_1 có hai phần tử trở lên, khi đó ta cần phải có một cách giải thích khác.
- Ở đây, ta sẽ đặt:

$$\rho_0(\theta) = \frac{\mathrm{p}(\theta)}{\pi_0} \text{ for } \theta \in \Theta_0 \quad \text{và} \quad \rho_1(\theta) = \frac{\mathrm{p}(\theta)}{\pi_1} \text{ for } \theta \in \Theta_1$$

Với $\mathrm{p}(\theta)$ là xác suất tiên nghiệm của θ.

- Như vậy, ta có thể hiểu rằng $\rho_0(\theta)$ là giá trị $\mathrm{p}(\theta)$ đẩy xuống không gian tham số Θ_0, tương tự như vậy với $\rho_1(\theta)$ và Θ_1.
- Từ đó, ta có:

$$\mathrm{p}_0 = \mathrm{P}(\theta \in \Theta_0|\mathrm{x}) = \int_{\theta \in \Theta_0} \mathrm{p}(\theta|\mathrm{x})\mathrm{d}\theta$$

- Sử dụng phép tỉ lệ thuận, ta được:

$$\int_{\theta \in \Theta_0} \mathrm{p}(\theta|\mathrm{x})\mathrm{d}\theta \sim \int_{\theta \in \Theta_0} \mathrm{p}(\theta).\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$

$$= \pi_0 \int_{\theta \in \Theta_0} \mathrm{p}(\mathrm{x}|\theta).\rho_0(\mathrm{x}|\theta)\mathrm{d}\theta$$

- Bằng một cách tương tự, ta cũng có một biểu thức cho p_1:

$$\mathrm{p}_1 \sim \pi_1 \int_{\theta \in \Theta_1} \mathrm{p}(\mathrm{x}|\theta).\rho_1(\mathrm{x}|\theta)\mathrm{d}\theta$$

Như vậy, ta đã tìm ra được công thức tính toán của hai đại lượng kiểm định ứng với phương pháp tiếp cận kiểu Bayes.

- Cuối cùng, ta sẽ tìm được đại lượng chênh lệch Bayes ưu tiên H_0 so với H_1 như sau:

$$\mathrm{B} = \frac{\int_{\theta \in \Theta_0} \mathrm{p}(\mathrm{x}|\theta).\rho_0(\mathrm{x}|\theta)\mathrm{d}\theta}{\int_{\theta \in \Theta_1} \mathrm{p}(\mathrm{x}|\theta).\rho_1(\mathrm{x}|\theta)\mathrm{d}\theta}$$

Một vài kết luận:

- Giá trị phân rã Bayes B dựa vào các đại lượng “likelihoods” như ρ_0, ρ_1 và $\mathrm{p}(\mathrm{x}|\theta)$ (hoặc là $\mathrm{p}(\mathrm{D}|\theta)$).
- Vì thế, giá trị phân rã Bayes B không thể là đại lượng đo lường sự khác biệt giữa các giả thuyết được suy ra trực tiếp từ dữ liệu có được.
- Tuy nhiên trong trường hợp đặc biệt, việc giới hạn số các giá trị ρ_0, ρ_1 có được sẽ khiến giá trị B trở nên tự nhiên hơn và sẽ là một giá trị có tính đánh giá cao.

Ví dụ:

Kiểm định giả thuyết một bên:

- Ta sẽ làm việc với những trường hợp cụ thể hơn cho bài toán kiểm định giả thuyết ở trên.
- Trong trường hợp ta xét:

$$\Theta = \{\text{Tập các biến giá trị thực}\}$$

- Khi đó các θ trong Θ cũng có giá trị thực.

- Ta xét các tập Θ_0 và Θ_1 và các giả thuyết $\mathrm{H}_0, \mathrm{H}_1$ như sau:

$$\mathrm{H}_0: \theta_0 < \theta_1 \text{ với các } \theta_0 \in \Theta_0 \text{ và } \theta_1 \in \Theta_1$$

$$\mathrm{H}_1: \theta_1 < \theta_0 \text{ với các } \theta_0 \in \Theta_0 \text{ và } \theta_1 \in \Theta_1$$

- Đối với những người theo trường phái Bayes, trường hợp trên không có gì đặc biệt, tuy nhiên nếu xét trong mối liên hệ với trường phái xác suất cổ điển, ta lại thấy sự liên hệ với giá trị p

Giả thuyết không dạng điểm:

- Trong xác suất truyền thống, ta có phép kiểm định dạng điểm quen thuộc và được sử dụng rộng rãi như sau:

$$\mathrm{H}_0: \theta = \theta_0$$

$$\mathrm{H}_1: \theta \neq \theta_0$$

- Ta sẽ sử dụng phương pháp Bayes để nghiên cứu trường hợp này.

- Để phù hợp với quan điểm Bayes, ta sẽ thay đổi bài toán như sau:

$$\mathrm{H}_0 : \theta \in \Theta_0 = (\theta_0 - \epsilon, \theta_0 + \epsilon)$$

$$\mathrm{H}_1 : \theta \in \Theta_1 =]\theta_0 - \epsilon, \theta_0 + \epsilon[$$

- Lí do cho việc lựa chọn:

❖Phù hợp với mô tả cho bài toán kiểm định giả thuyết theo trường phái Bayes ở phần trước.

❖Khi ϵ càng bé thì khoảng giá trị càng hẹp và ta có thể xấp xĩ θ về gần θ_0.

❖ Khi cho $\theta = \theta_0$, ta không có bất kì phương pháp tính toán trực tiếp nào cho hai đại lượng θ và θ_0. Ta không thể dùng các kĩ thuật tính toán thông thường.

Tính toán trong bài toán giả thuyết 0:

- Xét $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}$ là dãy các phép thử. Ta sẽ đưa ra từng đại lượng cần thiết cho việc tính toán.
- Giá trị tiên nghiệm π_0, π_1 với $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$:

$$\pi_0 = \int_{\theta_0 - \epsilon}^{\theta + \epsilon} \mathrm{p}(\theta)\mathrm{d}\theta$$

Và $\pi_1 = 1 - \pi_0$.

- Giá trị hậu nghiệm ρ_0, ρ_1:

$$\mathrm{p}_0 = \mathrm{P}(\theta \in \Theta_1|\mathrm{x}) = \pi_0 \int_{\theta \in \Theta_0} \mathrm{p}(\mathrm{x}|\theta).\ \rho_0(\theta)\mathrm{d}\theta$$

$$\mathrm{p}_1 = \mathrm{P}(\theta \in \Theta_1|\mathrm{x}) = \pi_1 \int_{\theta \in \Theta_1} \mathrm{p}(\mathrm{x}|\theta).\ \rho_1(\theta)\mathrm{d}\theta$$

- Chú ý:

$$\mathrm{p}(\mathrm{x}) = \pi_0 \mathrm{p}(\mathrm{x}|\theta_0) + \pi_1 \int \rho_1(\theta)\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$

$$= \pi_0 \mathrm{p}(\mathrm{x}|\theta_0) + \pi_1 \mathrm{p}_1(\mathrm{x})$$

- Biểu thức trên có được vì:

$$p(x) = \int_{-\infty}^{+\infty} p(x, \theta) d\theta$$

$$= \int_{-\infty}^{+\infty} p(\theta).\, p(x|\theta) d\theta \;\; (\text{định lý bayes})$$

- Bằng tính chất cơ bản của tích phân, ta có:

$$\int_{-\infty}^{+\infty} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$
$$= \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta + \int_{-\infty}^{\theta_0-\epsilon} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$
$$+ \int_{\theta_0+\epsilon}^{+\infty} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$

- Chú ý rằng:

$$\int_{\theta_0-\epsilon}^{\theta_0+\epsilon} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta = \mathrm{p}(\theta_0).\,\mathrm{p}(\mathrm{x}|\theta_0)$$

$$= \pi_0.\,\mathrm{p}(\mathrm{x}|\theta_0)$$

- Phần còn lại, ta sẽ được:

$$\int_{-\infty}^{\theta_0-\epsilon} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta + \int_{\theta_0+\epsilon}^{+\infty} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$

$$= \int_{\theta\in\Theta_1} \mathrm{p}(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta = \int_{\theta\in\Theta_1} \rho_1(\theta).\,\pi_1.\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$

$$= \pi_1.\int_{\theta\in\Theta_1} \rho_1(\theta).\,\mathrm{p}(\mathrm{x}|\theta)\mathrm{d}\theta$$

- Bằng định lý Bayes, ta có:

$$p_0 = p(\theta_0|x) = \frac{p(\theta_0).p(x|\theta_0)}{p(x)} = \frac{\pi_0.p(x|\theta_0)}{p(x)}$$

$$p_1 = \frac{\pi_1.\int \rho_1(\theta)p(x|\theta)d\theta}{p(x)} = \frac{\pi_1.p_1(x)}{p(x)}$$

- Như vậy, ta có thể tính được chỉ số độ lệch Bayes:

$$B = \frac{p_0 / p_1}{\pi_0 / \pi_1}$$

$$= \frac{p(x|\theta_0)}{p(x)}$$

Trường hợp cụ thể với phân phối chuẩn:

- Ta sẽ tìm hiểu bài toán trên trong trường hợp cụ thể với phân phối chuẩn.
- Giả sử ta có chuỗi phép thử $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}$ với $\mathrm{x}_\mathrm{i} \sim \mathrm{N}(\theta, \phi)$. Ta giả sử rằng ϕ đã được cố định.
- Ta sẽ lấy $\overline{\mathrm{x}} = \frac{\sum \mathrm{x}_\mathrm{i}}{\mathrm{n}} \sim \mathrm{N}\left(\theta, \frac{\phi}{\mathrm{n}}\right)$.
- Ta sẽ áp dụng phương pháp bayes.

- Để áp dụng phương pháp Bayes, ta cần lựa chọn phân phối tiên nghiệm của bài toán.
- Giả sử rằng: $\rho_1 \sim \mathrm{N}(\mu, \psi)$.
- Ta sẽ đặt $\theta_0 = \mu$.
- Như vậy bài toán lúc này đã có đủ các đại lượng: θ, θ_0, ρ_1 để áp dụng định lý vừa chứng minh.

- Trước hết, ta tính:

$$\mathrm{P}\left(\overline{\mathrm{x}'}, \theta_0\right) \text{ và } \mathrm{P}(\overline{\mathrm{x}'})$$

- Lưu ý rằng trong tình huống này, vì sau khi chọn được phân phối tiên nghiệm và đại lượng θ_0, $\overline{\mathrm{x}}$ sẽ trở thành một một bộ dữ liệu hoàn toàn mới, dựa trên dữ liệu cũ là $\overline{\mathrm{x}'}$.

- Ta biết:

$$\overline{\mathrm{x}'} = \overline{\mathrm{x}'} - \theta + \theta$$

Như là tổng của các biến ngẫu nhiên.

- $\overline{\mathrm{x}'} - \theta$ trừ đi chính trung bình của nó nên có dạng $\mathrm{N}\left(0, \frac{\phi}{\mathrm{n}}\right)$
- Ta được xác suất tiên nghiệm của $\theta \sim \mathrm{N}(\theta_0, \psi)$

- Như vậy ta có được:

$$\overline{\mathrm{x}'} \sim \mathrm{N}\left(\theta_0, \frac{\phi}{\mathrm{n}} + \psi\right)$$

- Ta cũng có:

$$\left(\overline{\mathrm{x}'} \middle| \theta\right) \sim \mathrm{N}\left(\theta_0, \frac{\phi}{\mathrm{n}}\right)$$

- Như vậy ta sẽ được:

$$B = \frac{\left\{\frac{2\pi\phi}{n}\right\}^{-\frac{1}{2}} \exp\left[-\frac{\frac{1}{2}(\bar{x} - \theta_0)^2}{\frac{\phi}{n}}\right]}{\left\{2\pi\left(\psi + \frac{\phi}{n}\right)\right\}^{-\frac{1}{2}} \exp\left[-\frac{\frac{1}{2}(\bar{x} - \theta_0)^2}{\psi + \frac{\phi}{n}}\right]}$$

- Đặt:

$$z = \frac{|\bar{x} - \theta_0|}{\sqrt{\frac{\phi}{n}}}$$

- Lúc này ta sẽ thay z vào biểu thức trên.

- Ta sẽ được kết quả:

$$\mathrm{B} = \left\{1 + \frac{\mathrm{n}\psi}{\phi}\right\}^{\frac{1}{2}} . \exp\left[-\frac{1}{2}\mathrm{z}^2 \left\{1 + \frac{\phi}{\mathrm{n}\psi}\right\}^{-1}\right]$$

Bài tập cụ thể:

- Giả sử ta có chuỗi phép thử $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}$ với $\mathrm{x}_\mathrm{i} \sim \mathrm{N}(\theta, 1)$.
- Giả sử $\theta \sim \mathrm{N}(\theta_0, 5)$
- Ta có hai giả thuyết sau:

$$\mathrm{H}_0 : \theta \in \Theta_0 = (\theta_0 - \epsilon, \theta_0 + \epsilon)$$

$$\mathrm{H}_1 : \theta \in \Theta_1 =]\theta_0 - \epsilon, \theta_0 + \epsilon[$$

- Hãy kiểm định xem giả thuyết nào có vẻ đúng nhất ?