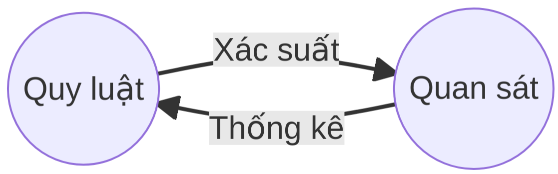
XSTK

Xác Suất Thống Kê

Võ Chí Công

Table of contents

Lời nói đầu



Hình 0.1: Xác suất và thống kê.

Xác suất thống kê là nền tảng giúp ích cho việc phân tích dữ liệu, tổ chức thực hành thí nghiệm, chạy các mô hình giả lập, giải các bài toán tìm nghiệm tối ưu, nghiên cứu và ứng dụng học máy.

Môn xác suất nghiên cứu cách suy luận ra các đặc tính của tập dữ liệu sẽ được tạo ra từ một nguyên lý, quy trình sản sinh dữ liệu. Ngược lại, môn thống kê nghiên cứu cách dự đoán đặc tính của một quy trình sản sinh dữ liệu từ tập dữ liệu về hiện tượng đã phát sinh và được quan sát. H 0.1 minh hoạ quan hệ giữa "xác suất" và "thống kê".

Phân tích, khai thác dữ liệu, học máy và khoa học dữ liệu là những tên gọi khác của thống kê, tùy theo bối cảnh và trào lưu. Một số ứng dụng cụ thể của thống kê là tính toán hồi quy, mật độ, phân loại và giả lập.

Động cơ

Môn "xác suất" tôi đã học những kiến thức cơ bản vài lần, tính ra là ở cấp 3, trong đại học, và gần đây là học trực tuyến. Môn "thống kê" tôi chưa học được cho ra bài bản lần nào, mấy tháng đầu năm 2022 có thử sức học trực tuyến nhưng đầu tư thời gian không đủ nên thi rớt thảm hại.

Lần này nhất quyết học lại môn thống kê một cách nghiêm túc hơn, tôi tóm tắt lại kiến thức xác suất thống kê bằng tiếng Việt, mặc dù tài liệu học hầu hết là tiếng Anh, tiếng Nhật. Hy vọng tiếng mẹ đẻ sẽ giúp tôi hiểu rõ hơn các vấn đề, và trau dồi vốn từ vựng để chia sẻ kiến thức với các đồng nghiệp và bạn bè người Việt.

Động cơ của việc học xác suất thống kê của tôi là để hiểu rõ hơn các lý thuyết căn bản trong ngành học máy và trí tuệ nhân tạo và áp dụng vào thực tiễn một cách đúng đắn, an toàn và công bằng hơn.

Mục tiêu cụ thể trước mắt tôi đặt ra là học hiểu và lấy được chứng chỉ hoàn thành khoá học Fundamentals of Statistics của Philippe Rigollet (2022), giáo sư đại học MIT dạy trên nền tảng trực tuyến edX. Tài liệu tham khảo là quyển "All of Statistics" của Wasserman (2004), cũng chính là tài liệu tham khảo của khoá học nêu trên.

Chú ý

Tài liệu này không đi sâu vào các chứng minh chi tiết, nhưng sẽ cố gắng ghi rõ các công thức và định nghĩa. Thuật ngữ chuyên môn trong tài liệu này chắc chắn có nhiều chỗ không chuẩn chỉnh do vốn tiếng Việt và kiến thức hạn chế của tác giả. Xin vui lòng góp ý tại GitHub issues.

Tài liệu này được viết bằng các công cụ là Quarto và VSCode. Truy cập trực tuyến tại xstk .

Ký hiệu

Ký hiệu	Ý nghĩa
$\mathbb{R}$	Tập hợp số thực
$\mathbb{R}_+$	Tập hợp số thực dương
$f(\alpha) \propto g(\alpha)$	f/g không phụ thuộc α
$\mathcal{N}\left(\mu, \sigma^2\right)$	phân phối chuẩn tâm μ, lệch $\lvert\sigma\rvert$

Phần I

Xác suất

1 Xác suất

1.1 Sự kiện

Không gian Ω là tập hợp chứa tất cả những **hiện tượng** ω có thể xảy ra từ một **thí nghiệm**. Các tập con của Ω là các **sự kiện**.

Ví dụ xem xét thí nghiệm tung một đồng xu đúng hai lần, quan sát đồng xu rớt xuống nằm ngửa (H) hay sấp (T), ta có $\Omega = \{HH, HT, TH, TT\}$ bao gồm 4 kết quả có thể xảy ra. Sự kiện lần tung đầu tiên ra mặt ngửa của đồng xu là tập hợp $\{HH, HT\}$.

Cho một sự kiện $A \subseteq \Omega$, ta nói A **xảy ra**, hoặc A là **đúng**, nếu có một hiện tượng $\omega \in A$ **xảy ra**. Sự kiện **ngược lại** với A là $A^c := \Omega - A := \{\omega \in \Omega : \omega \notin A\}$, tức là “không xảy ra A”. Theo định nghĩa, rõ ràng Ω luôn luôn đúng, còn sự kiện rỗng $\emptyset \equiv \Omega^c$ luôn luôn sai. Cho thêm sự kiện B, ta có $A \cup B$ là sự kiện “A **hoặc** B ít nhất một việc xảy ra”, còn $AB := A \cap B$ là sự kiện “A **và** B đồng thời xảy ra”.

Chuỗi sự kiện $A_1, A_2, \ldots$ được gọi là **phân ly** nếu $A_iA_j \equiv \emptyset$ với mọi $i \neq j$. Khi đó nếu $A_1 \cup A_2 \cup \cdots \equiv C$ thì ta nói $A_1, A_2, \ldots$ là một cách **phân hoạch** sự kiện C.

1.2 Xác suất

Nếu một xạ ảnh $\mathbb{P}$ từ không gian các sự kiện $A \subseteq \Omega$ lên tập hợp số thực $\mathbb{R}$ thỏa mãn các điều kiện:

1. $\mathbb{P}(A) \geq 0 \quad \forall A$
2. $\mathbb{P}(\Omega) = 1$
3. Nếu chuỗi $A_1, A_2, \ldots$ phân hoạch C thì $\mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \mathbb{P}(C)$

thì ta gọi $\mathbb{P}$ là một **phân phối xác suất** hoặc **độ đo xác suất**.

Có hai cách cắt nghĩa khái niệm xác suất là tần suất và niềm tin. Theo cách hiểu tần suất thì $\mathbb{P}(A)$ chính là tỷ lệ số lần sự kiện A xảy ra nếu ta thực hiện thí nghiệm vô số lần. Còn theo cách hiểu niềm tin thì $\mathbb{P}(A)$ là thước đo mức độ mà một quan sát viên tin tưởng rằng hiện tượng A sẽ xảy ra.

1.3 Tham số

Định nghĩa 1.1 (Biến ngẫu nhiên, random variable)**.** là một quy tắc ánh xạ $X : \Omega \to \mathbb{R}$ gán cho mỗi hiện tượng ω trong không gian Ω một số thực $X(\omega)$.

Định nghĩa 1.2 (Hàm phân phối tích lũy, cumulative distribution function)**.** là hàm tăng đơn điệu và liên tục bên phải $F : \mathbb{R} \to [0,1], F(x) \coloneqq \mathbb{P}(X \leq x)$ với biến $X : \Omega \to \mathbb{R}$.

Định nghĩa 1.3 (Phân vị)**.** Biến X tuân theo phân phối tích lũy F có phân vị mức $1-\alpha$ là $q_\alpha \coloneqq F^{-1}(1-\alpha)$.

Định nghĩa 1.4 (Quartile, median, mode)**.** Ta gọi $q_{3/4}, q_{1/2}, q_{1/4}$ là tứ phân vị đầu tiên (first quartile), trung vị (median), và tứ phân vị thứ ba. Gọi mode là giá trị biến số có xác suất cao nhất.

Định nghĩa 1.5 (Hàm khối xác suất, probability mass function)**.** Với biến rời rạc $X : \Omega \to D$, $D = \{x_1, x_2, \ldots\}$ đếm được, thì ta có thể gán xác suất $\mathbb{P}(x_i) = p_i \geq 0, \forall i = 1, 2, \ldots$ sao cho $\sum_i p_i \equiv 1$.

Định nghĩa 1.6 (Hàm mật độ xác suất, probability density function)**.** là $f(x) = F'(x)$ với $F(x) \coloneqq \mathbb{P}(X \leq x)$ là phân phối tích lũy của biến $X : \Omega \to \mathbb{R}$. Mật độ xác suất là không âm và có tích phân toàn phần bằng 1.

Định nghĩa 1.7 (PDF of bivariate distributions)**.** Với hai biến liên tục $(X, Y) \in \mathbb{R}^2$ ta gọi hàm không âm $f(x, y)$ là mật độ xác suất đồng thời nếu:

$$\int\int_A f(x, y)dxdy \equiv \mathbb{P}((X, Y) \in A), \quad \forall A \subset \mathbb{R}^2.$$

Mật độ xác suất biên và mật độ xác suất có điều kiện tương ứng là:

$$f(x) := \int f(x, y)dy,$$
$$f(y|x) := \frac{f(x, y)}{f(x)}.$$

Định lý 1.1 (Chuẩn hóa mật độ xác suất)**.** *Nếu có độ đo $g(x) \geq 0, x \in \mathbb{R}$ khả tích toàn phần thì có mật độ xác suất*

$$f(x) := \frac{g(x)}{\int_{\mathbb{R}} g(x)dx} \propto g(x).$$

Định lý 1.2 (Biến đổi tham số)**.** *Giả sử X là tham số có mật độ xác suất f_X, hàm $g : \mathbb{R} \to \mathbb{R}$ có vi phân > 0. Khi đó tham số $Y = g(X)$ có mật độ xác suất là*

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{g'(g^{-1}(y))} \quad \forall y \in g(\mathbb{R}).$$

Chứng minh. Xem "4.1.3 Functions of Continuous Random Variables", Pishro-Nik (2014).

□

1.4 Xác suất đồng thời

Định nghĩa 1.8 (Xác suất hợp)**.** Ký hiệu $\mathbb{P}(A, B)$ hoặc $\mathbb{P}(A \cap B)$ chỉ xác suất sự kiện A và sự kiện B đồng thời xảy ra.

Định nghĩa 1.9 (Độc lập)**.** Hai sự kiện A và B là độc lập nếu $\mathbb{P}(A, B) \equiv \mathbb{P}(A)\mathbb{P}(B)$.

Hai biến X và Y là độc lập nếu hai sự kiện $X \leq x$ và $Y \leq y$ là độc lập đối với mọi x, y.

Định nghĩa 1.10 (IID)**.** Các biến ngẫu nhiên $X_1, X_2, .$ được gọi là iid ("independent and identically distributed", "độc lập và phân phối giống nhau") nếu chúng cùng tuân theo duy nhất (identical) một phân phối xác suất, và từng cặp biến là độc lập (independent) với nhau. Dùng biến X để thể hiện phân phối xác suất chung đó, ta viết

$$X_1, \dots, X_n \overset{iid}{\sim} X.$$

Định nghĩa 1.11 (Xác suất có điều kiện)**.** Ký hiệu $\mathbb{P}(A|B)$ hoặc $\mathbb{P}_B(A)$ chỉ xác suất của sự kiện A, khi biết sự kiện B đã xảy ra,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}.$$

Định lý 1.3 (Định lý Bayes)**.**

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}.$$

1.5 Đặc trưng

Định nghĩa 1.12 (Trung bình)**. Trung bình**, hay **giá trị kỳ vọng** của biến X là

$$\mathbb{E}\left[X\right] := \int xdF(x) := \begin{cases} \sum_x xp(x) & \text{if } X \text{ is discrete} \\ \int xf(x)dx & \text{if } X \text{ is continuous.} \end{cases}$$

Nếu giá trị $\int |x|dF(x) < \infty$ ta nói là $\mathbb{E}\left[X\right]$ "tồn tại" (well-defined).

Ta có

$$\mathbb{E}\left[X+Y\right] \equiv \mathbb{E}\left[X\right] + E[Y].$$

Định nghĩa 1.13 (Sample mean)**.** Cho mẫu $X_1, \ldots, X_n \overset{iid}{\sim} X$. Trung bình mẫu là

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Định nghĩa 1.14 (Hiệp phương sai)**.**

$$\begin{aligned} \operatorname{Cov}\left(X,Y\right) &:= \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right] \\ &\equiv \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \end{aligned}$$

Nếu X và Y độc lập thì $\operatorname{Cov}\left(X,Y\right) = 0$.

Định nghĩa 1.15 (Phương sai)**.** Phương sai, hay phân tán (variance) là

$$\operatorname{Var}\left(X\right) := \operatorname{Cov}\left(X,X\right) \equiv \mathbb{E}\left[X^2\right] - (\mathbb{E}\left[X\right])^2.$$

Remark.

$$\operatorname{Var}\left(X+Y\right) \equiv \operatorname{Var}\left(X\right) + \operatorname{Var}\left(Y\right) + 2\operatorname{Cov}\left(X,Y\right).$$

Định nghĩa 1.16 (Unbiased sample variance)**.** Cho mẫu $X_1, \ldots, X_n \overset{iid}{\sim} X$. Phương sai mẫu không lệch (ĐN 4.2) là

$$S_n^2 := \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}.$$

Ta có $\mathbb{E}\left[S_n^2\right] \equiv \mathrm{Var}\left(X\right).$

Remark. Định lý Cochoran ĐL 6.1.

Định lý 1.4 (Iterated Expectations)**.** *Với các biến X, Y, giả sử các giá trị kỳ vọng tồn tại, ta có*

$$\mathbb{E}\left[\mathbb{E}\left[Y|X\right]\right] \equiv \mathbb{E}\left[Y\right].$$

1.6 Tích suất

Định nghĩa 1.17 (Tích suất)**.** Tích suất (moment) thể hiện trọng tâm, độ phân tán, hay độ lệch của phân phối. Tích suất bậc n của biến X với mật độ $f(x)$ là:

$$\mathbb{E}\left[X^n\right] := \int x^n f(x) dx$$

Định nghĩa 1.18 (Hàm tạo tích suất)**.** "Moment generation function (MGF)" của biến X là

$$\psi_X(t) := \mathbb{E}\left[e^{tX}\right], t \in \mathbb{R}.$$

Nếu MGF "tồn tại" lân cận 0 thì đạo hàm cấp k của ψ tại 0 chính là:

$$\psi_X^{(k)}(0) \equiv \mathbb{E}\left[X^k\right].$$

Định nghĩa 1.19 (Tích suất mẫu)**.** Từ các quan sát $X_1, \ldots X_n \overset{iid}{\sim} \mathbb{P}_\theta, \theta \in \Theta \subset \mathbb{R}^d$. Để ước lượng tích suất $\mathbb{E}\left[X^k\right]$ ta có thể dùng các *sample moment*:

$$\hat{m}_k := \frac{1}{n}\sum_{i=1}^n X_i^k, \quad k = 1, 2, \ldots.$$

2 Phân phối

Có một số phân phối xác suất thông dụng.

2.1 Phân phối Bernoulli

Định nghĩa 2.1 (Bernoulli distribution)**.** $X \sim \text{Ber}(p)$ có

$$\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$$

và $\mathbb{E}[X] = p, \text{Var}(X) = p(1 - p)$.

2.2 Phân phối nhị thức

Định nghĩa 2.2 (Binomial distribution)**.** $K \sim \text{Binom}(n, p)$ với $n \in \mathbb{Z}_+, p \in (0, 1)$ mô tả tổng số lần thành công của n thí nghiệm độc lập $K_1, \dots, K_n \overset{iid}{\sim} \text{Ber}(p)$. Ta có

$$\mathbb{P}(K = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

và $\mathbb{E}[K] = np, \text{Var}(K) = np(1 - p)$.

2.3 Phân phối Poisson

Định nghĩa 2.3 (Phân phối Poisson). $K \sim \text{Poiss}(\lambda)$, $\lambda > 0$ cố định, có

$$\mathbb{P}(K = k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

$k = 0, 1, 2, \ldots$ là số lần phát sinh sự kiện trong một giới hạn cố định, các sự kiện phát sinh độc lập.

$$\mathbb{E}\left[K\right] = \text{Var}\left(K\right) = \lambda.$$

Khi n đủ lớn và p đủ nhỏ thì $\text{Poiss}(np)$ gần với $\text{Binom}(n, p)$.

2.4 Phân phối hình học

Định nghĩa 2.4 (Geometric distribution). $K \sim \text{Geom}(p)$, $p \in (0, 1)$ có

$$\mathbb{P}(K = k) = p(1-p)^{k-1}, k = 1, 2, \ldots$$

là xác suất thành công đầu tiên xảy ra sau đúng k lần thực nghiệm $\text{Ber}(k)$.

$$\mathbb{E}\left[K\right] = 1/p, \text{Var}\left(K\right) = (1-p)/p^2.$$

2.5 Phân phối liên tục

2.6 Phân phối đều

Định nghĩa 2.5 (Uniform distribution). $X \sim \text{Unif}[a, b]$ có mật độ

$$f(x) := \frac{x-a}{b-a} \text{ for } x \in [a, b]$$

và

$$\mathbb{E}\left[X\right] = \frac{a+b}{2}, \text{Var}\left(X\right) = \frac{(b-a)^2}{12}.$$

Nếu biến X có cdf F (ĐN 1.2) khả nghịch thì $F(X) \sim \text{Unif}[0, 1]$.

2.7 Phân phối mũ

Định nghĩa 2.6 (Exponential distribution)**.** $X \sim \operatorname{Exp}(\lambda)$ có

$$f(x) = \lambda e^{-\lambda x}, x \in \mathbb{R}_+,$$

x ước lượng khoảng cách giữa hai lần phát sinh sự kiện trong quá trình $\operatorname{Poiss}(\lambda)$.

$$\mathbb{E}\left[X\right] = 1/\lambda, \operatorname{Var}\left(X\right) = 1/\lambda^2.$$

2.8 Phân phối chuẩn

Định nghĩa 2.7 (Gaussian Distribution)**.** $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ có mật độ

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

và $\mathbb{E}\left[X\right] = \mu, \operatorname{Var}\left(X\right) = \sigma^2$.

Định lý 2.1 (Tổng của các phân phối chuẩn)**.** *Nếu* $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$ $(i = 1, \ldots, n)$ *thì*

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Định nghĩa 2.8 (Vector phân phối chuẩn)**.** Nếu mọi tổ hợp tuyến tính các yếu tố của vector $\boldsymbol{v}$ đều thuộc phân phối chuẩn 1 biến thì $\boldsymbol{v}$ được gọi là một vector phân phối chuẩn.

2.9 Phân phối χ^2

Định nghĩa 2.9 (Phân phối χ^2)**.** $X \sim \chi_k^2$ là tổng bình phương của $X_1, \ldots, X_k \overset{iid}{\sim} \mathcal{N}(0,1)$

$$X := \sum_{i=0}^{k} X_i^2$$

và có $\mathbb{E}[X] = k, \mathrm{Var}(X) = 2k$.

2.10 Phân phối Student

Định nghĩa 2.10 (Student's T distribution)**.** Nếu có $Z \sim \mathcal{N}(0,1), V \sim \chi_k^2$ độc lập với nhau thì

$$X := \frac{Z}{\sqrt{V/k}}$$

tuân theo phân phối Student's t_k.

2.11 Phân phối Beta

Định nghĩa 2.11 (Beta distribution)**.** $X \sim \mathrm{B}\alpha\beta$, tham số $\alpha > 0, \beta > 0$ có mật độ trên $x \in [0,1]$ là

$$f(x) = Cx^{\alpha-1}(1-x)^{\beta-1},$$

$C = C(\alpha, \beta)$ là hằng số chuẩn hóa.

Phân phối Beta có đồ thị rất linh hoạt trên khoảng $[0,1]$, rất phù hợp để sử dụng làm xác suất tiền nghiệm (ĐN 7.1) cho một tham số xác suất $p \in [0,1]$.

Ví dụ Unif$[0,1] = \mathrm{B}11$.

2.12 Phân phối Gamma

Định nghĩa 2.12 (Gamma distribution)**.** với tham số $q > 0, \lambda > 0$, có mật độ là

$$f(x) := \frac{\lambda^q x^{q-1} e^{-\lambda x}}{\Gamma(q)} \quad \forall x \in (0, \infty),$$

Γ là hàm Euler Gamma.

3 Hội tụ

Có một số kiểu hội tụ của biến ngẫu nhiên.

3.1 Hội tụ xác suất

Cho một chuỗi biến ngẫu nhiên $X_1, X_2, .$ và một biến ngẫu nhiên X.

Định nghĩa 3.1 (Hội tụ gần tuyệt đối).

$$X_n \xrightarrow[n\to\infty]{\text{a.s.}} X \iff \mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1.$$

Định nghĩa 3.2 (Hội tụ theo xác suất).

$$X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \iff \mathbb{P}(|X_n - X| > \epsilon) \xrightarrow[n\to\infty]{} 0, \quad \forall \epsilon > 0.$$

Định nghĩa 3.3 (Hội tụ theo phân phối).

$$X_n \xrightarrow[n\to\infty]{(d)} X \iff \mathbb{P}[X_n(x) \leq x] \xrightarrow[n\to\infty]{} \mathbb{P}[X \leq x]$$

tại mọi điểm x mà cdf của X liên tục.

Định nghĩa 3.4 (Hội tụ về chuẩn). (asymptotically normal) với phương sai σ^2

$$\frac{\sqrt{n}}{|\sigma|}\left(X_n - X\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)\,.$$

3.2 Độ mạnh

Định lý 3.1. *Xếp theo thứ tự từ mạnh đến yếu:*

$$X_n \xrightarrow[n\to\infty]{a.s.} X \implies X_n \xrightarrow[n\to\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to\infty]{(d)} X.$$

Nếu $X_n \xrightarrow[n\to\infty]{(d)} X$, *và* X *có mật độ xác suất, thì* $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$.

Nếu chuỗi X_n *có* $\mathbb{E}[X_n] \to \mu$ *và* $Var(X_n) \to 0$ *thì* $X_n \xrightarrow{\mathbb{P}} \mu$.

Nếu $X_n \xrightarrow[n\to\infty]{\mathbb{P}} X$ *thì* $\mathbb{P}(a \le X_n \le b) \xrightarrow[n\to\infty]{} \mathbb{P}(a \le X \le b)$ *với mọi khoảng* $[a,b]$.

3.3 Tổng và tích

Định lý 3.2. *Nếu có hai chuỗi biến ngẫu nhiên* X_n, Y_n *hội tụ gần tuyệt đối hoặc hội tụ theo xác suất về* X, Y, *thì tổng* $X_n + Y_n$ *và tích* $X_n Y_n$ *cũng hội tụ tương tự (gần tuyệt đối, hoặc theo xác suất) về tổng* $X + Y$ *và tích* XY.

3.4 Slutsky

Định lý 3.3. *Hơn nữa, ở ĐL 3.2 nếu* $Y_n \xrightarrow{\mathbb{P}} y$, y *là một số thực cố định thì có thể nới lỏng điều kiện đối với* X_n *thành hội tụ theo phân phối.*

3.5 Ánh xạ liên tục

Định lý 3.4 (Continuous mapping)**.** *Nếu* $X_n \xrightarrow[n\to\infty]{a.s./\mathbb{P}/(d)} X$ *thì đối với mọi hàm* f *liên tục:*

- $f(X_n) \xrightarrow[n\to\infty]{a.s./\mathbb{P}/(d)} f(X)$.
- $\mathbb{E}\left[f(X_n)\right] \xrightarrow[n\to\infty]{} \mathbb{E}\left[f(X)\right]$ *nếu* f *còn bị chặn.*

3.6 Đại đa số

Định lý 3.5 (Law of Large Numbers, LLN)**.** *Cho* n *biến iid* $X_1, X_2, .., X_n \overset{iid}{\sim} X$ *có* $\mathbb{E}\left[X\right] < \infty$*. Khi đó trung bình mẫu (ĐN 1.13) hội tụ:*

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P},\, a.s.} \mathbb{E}\left[X\right].$$

3.7 Hội tụ trung tâm

Định lý 3.6 (Central Limit Theorem, CLT)**.** *Giả sử thêm là* $Var(X_i) = \sigma^2 < \infty$*. Khi đó*

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{|\sigma|}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

3.8 Bất đẳng thức Hoefding

Định lý 3.7. *Nếu có một khoảng cố định* $[a, b]$ *gần như tuyệt đối (almost surely) chứa các biến* $X_i (i = 1, 2, \dots, n)$ *thì*

$$\mathbb{P}[|\bar{X}_n - \mu| \geq \epsilon] \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}, \quad \forall \epsilon > 0.$$

3.9 Phương pháp Delta

Định lý 3.8 (Phương pháp Delta)**.** *Giả sử chuỗi $(\theta_n)_{n\geq 1}$ chuẩn tiến (ĐN 3.4) với phương sai σ^2 về một điểm $\theta \in \mathbb{R}$. Giả sử $g : \mathbb{R} \to \mathbb{R}$ có vi phân g' liên tục, $\neq 0$ tại θ. Khi đó,*

$$\frac{\sqrt{n}}{|\sigma|}\left(\frac{g(\theta_n) - g(\theta)}{g'(\theta)}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

Định lý 3.9 (Phương pháp Delta nhiều biến)**.** *Giả sử chuỗi $(\boldsymbol{\theta}_n)_{n\geq 1}$ chuẩn tiến với phương sai $\Sigma(\boldsymbol{\theta})$ về $\boldsymbol{\theta} \in \mathbb{R}^d$:*

$$\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d\left(\mathbf{0}, \Sigma\right).$$

Giả sử $g : \mathbb{R}^d \to \mathbb{R}^k$ có vi phân ∇g liên tục $(k < d)$. Khi đó,

$$\sqrt{n}\left(g(\boldsymbol{\theta}_n) - g(\boldsymbol{\theta})\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\left(\mathbf{0}, \Gamma\right),$$

với $\Gamma := \nabla g(\boldsymbol{\theta})^T \Sigma \nabla g(\boldsymbol{\theta})$. Nếu Σ khả nghịch, ∇g rank k thì Γ khả nghịch,

$$\sqrt{n}\Gamma^{-1/2}\left(g(\boldsymbol{\theta}_n) - g(\boldsymbol{\theta})\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k\left(\mathbf{0}, \boldsymbol{I}_k\right),$$

$$n\Gamma^{-1}\left(g(\boldsymbol{\theta}_n) - g(\boldsymbol{\theta})\right)^2 \xrightarrow[n\to\infty]{(d)} \chi^2_k.$$

Phần II

Thống kê

4 Thống kê

4.1 Mô hình

Trên không gian đo được $E \subseteq \mathbb{R}$ ta quan sát các mẫu $X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$. Với tập tham số Θ ta xây dựng bộ độ đo xác suất $(\mathbb{P}_\theta)_{\theta\in\Theta}$ để mô phỏng $\mathbb{P}$.

Nếu tồn tại $\theta \in \Theta$ để $\mathbb{P}_\theta \equiv \mathbb{P}$ thì ta nói mô hình "hợp lệ" (well specified).

Nếu từ tập Θ, ánh xạ $\theta \mapsto \mathbb{P}_\theta$ là đơn ánh thì ta nói θ (hoặc $\mathbb{P}_\theta$) là "có thể xác định" (identifiable).

Nếu $\Theta \subseteq \mathbb{R}^d$ thì mô hình được nói có số biến hữu hạn (parametric model).

4.2 Ước lượng

Thống kê lượng (statistic) θ là bất cứ hàm số nào có thể đo được (measurable function) trên tập dữ liệu đối tượng X_i.

Đánh giá (estimator) θ_n đối với mục tiêu thống kê θ là một thống kê khác không phụ thuộc vào θ.

Đánh giá được gọi là nhất quán (consistent) nếu nó hội tụ về mục tiêu (ĐN 3.2).

Định nghĩa 4.1 (Pivotal statistic)**.** Một phân phối hay một ước lượng được gọi là pivotal nếu nó không phụ thuộc vào phân phối chưa biết và giá trị cụ thể của dữ liệu.

4.3 Sai số

Định nghĩa 4.2 (Độ lệch, bias)**.** giữa đánh giá θ_n với mục tiêu θ là

$$\text{bias}_\theta(\theta_n) := \mathbb{E}\left[\theta_n\right] - \theta.$$

Nếu độ lệch bằng 0 ta nói đánh giá không lệch (unbiased).

Định nghĩa 4.3 (Sai số bậc 2, quadratic risk, mean squared error)**.** giữa đánh giá θ_n với mục tiêu θ là

$$\text{MSE}(\theta_n) := \mathbb{E}\left[(\theta_n - \theta)^2\right] \equiv \text{bias}_\theta(\theta_n)^2 + \text{Var}\left(\theta_n\right)$$

bao hàm cả độ lệch và độ nhiễu của đánh giá.

4.4 Khoảng tin cậy

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$ xây dựng dựa trên quan sát $X_1, \dots, X_n$, giả sử số $\alpha \in (0, 1)$. Khoảng tin cậy (confidence interval) cấp $1-\alpha$ đối với θ là một khoảng ngẫu nhiên $\mathcal{I}$ (phụ thuộc vào $X_1, \dots, X_n$, không phụ thuộc vào θ), mà xác suất $\mathcal{I}$ có chứa θ là đủ cao:

$$\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall\theta \in \Theta.$$

Nếu

$$\lim_{n\to\infty} \mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha, \quad \forall\theta \in \Theta$$

thì ta gọi $\mathcal{I}$ là khoảng tin cậy tiệm cận cấp $1 - \alpha$ đối với θ.

4.5 Định lý cơ bản

Định nghĩa 4.4 (Empirical cdf)**.** Với $X_1, \dots, X_n \overset{iid}{\sim} X$, ước lượng phân phối tích lũy (ĐN 1.2) của X là hàm tăng đơn điệu và liên tục bên phải

$$F_n : \mathbb{R} \to [0, 1], F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\},$$

tức là tỷ lệ số biến $X_i \leq t$.

Nếu xếp $X_1 \le \ldots \le X_n < X_{n+1} = \infty$ từ nhỏ đến lớn thì ta có

$$F(t) = \frac{i}{n} \text{ for } t \in [X_i, X_{i+1}) \quad \forall i = 1, \ldots, n.$$

Áp dụng LLN lên biến $\mathbb{1}\{X < t\}$ ta có $F_n(t) \xrightarrow[n\to\infty]{\text{a.s.}} F(t)\, \forall t \in \mathbb{R}$.

Định lý 4.1 (Glivenko-Cantelli, Fundamental Theorem of Statistics)**.** *F_n hội tụ đồng đều (converge uniformly) lên F.*

$$\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Nói cách khác $\forall\delta > 0, \forall\epsilon > 0 \;:\; \exists N \;:$

$$n > N \implies \mathbb{P}\left(\sup_{t\in\mathbb{R}} |F_n(t) - F(t)| < \delta\right) \ge 1 - \epsilon.$$

Định lý 4.2 (Donsker's theorem)**.** *Nếu cdf F thật là liên tục thì*

$$\sqrt{n} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} Z \sim \sup_{0\le x\le 1} |B(x)|,$$

*với Z là phân phối của đỉnh giá trị tuyệt đối của **Brownian bridge**, và là phân phối pivotal(ĐN 4.1).*

5 Ước lượng

Chiến lược ước lượng phân phối *thật* (true distribution)

- Định nghĩa khoảng cách giữa các phân phối (TV distance)
- Ước lượng khoảng cách nêu trên (KL divergence)
- Tìm điểm cực tiểu của ước lượng nêu trên (minimization).

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ xây dựng dựa trên quan sát iid rv $X_1, \ldots, X_n$ trên tập mẫu E và bộ tham số Θ . Ngầm định tồn tại *tham số thật* $\theta^* \in \Theta$ để $X_1 \sim \mathbb{P}_{\theta^*}$.

5.1 Tổng biến động

5.1.1 Khoảng cách

Định nghĩa 5.1. tổng biến động (total variation distance) giữa hai độ đo xác suất $\mathbb{P}_\theta$ và $\mathbb{P}_\eta$ là

$$\mathrm{TV}(\mathbb{P}_\theta, \mathbb{P}_\eta) = \max_{A \subset E} | \, \mathbb{P}_\theta(A) - \mathbb{P}_\eta(A) \, |.$$

Định lý 5.1 (Công thức tính)**.** *Nếu tập mẫu E là rời rạc (discrete: countable or finite), xác suất $\mathbb{P}_\theta, \mathbb{P}_\eta$ có hàm khối lần lượt là p_θ, p_η thì*

$$TV(\mathbb{P}_\theta, \mathbb{P}_\eta) = \frac{1}{2} \sum_{x \in E} | \, p_\theta(x) - p_\eta(x) \, | \, .$$

Nếu tập mẫu E là liên tục (continuous), xác suất $\mathbb{P}_\theta, \mathbb{P}_\eta$ có mật độ lần lượt là f_θ, f_η thì

$$TV(\mathbb{P}_\theta, \mathbb{P}_\eta) = \frac{1}{2} \int_E | \, f_\theta(x) - f_\eta(x) \, | \, dx.$$

5.2 Phân kỳ KL

5.2.1 Phân kỳ KL

Định nghĩa 5.2. (KL divergence) Ký hiệu f là mật độ hoặc hàm khối xác suất:

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_\eta) \coloneqq \mathbb{E}_\theta \left[\ln \frac{f_\theta(x)}{f_\eta(x)} \right] = \mathbb{E}_\theta \left[\ln f_\theta(x) \right] - \mathbb{E}_\theta \left[\ln f_\eta(x) \right].$$

5.2.2 Đặc điểm

Proposition 5.1. *Phân kỳ KL thỏa mãn 2/4 đặc điểm của "khoảng cách":*

1. $KL(\mathbb{P}_\theta, \mathbb{P}_\eta) \geq 0$
2. $KL(\mathbb{P}_\theta, \mathbb{P}_\eta) \equiv 0 \iff \mathbb{P}_\theta \equiv \mathbb{P}_\eta$.

5.3 Hợp lý cực đại

Định nghĩa 5.3 (Hợp lý, Likelihood)**.**

$$L_\theta(x_1, \ldots, x_n) \coloneqq \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_\theta(X_i = x_i).$$

Định nghĩa 5.4 (Log likelihood)**.**

$$\ell_\theta(x_1, \ldots, x_n) \coloneqq \ln L_\theta(x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln p_\theta(X_i = x_i).$$

Định nghĩa 5.5 (Maximum Likelihood Estimator, MLE).

$$\hat{\theta}_n := \underset{\theta}{\operatorname{argmax}}\, L_\theta(x_1, \dots, x_n) \equiv \underset{\theta}{\operatorname{argmax}}\, \ell_\theta(x_1, \dots, x_n).$$

Định nghĩa 5.6 (Fisher information). Giả sử log likelihood đối với một quan sát X theo mô hình $\theta \in \mathbb{R}$ là $\ell = \ell(\theta) := \ln L_1(X, \theta), \theta \in \Theta \subset \mathbb{R}$. Giả sử $\ell(\theta)$ có đạo hàm bậc hai. Dưới một số điều kiện chuẩn, **thông tin Fisher** của mô hình được định nghĩa là

$$I(\theta) := \operatorname{Var}\left(\frac{\partial \ell}{\partial \theta}\right) = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right].$$

Trường hợp mô hình đa biến, $\theta := (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$:

$$H_{ij} := \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j},$$

$$I(\theta) := -\mathbb{E}\left[\begin{pmatrix} H_{11} & \cdots & H_{1k} \\ \vdots & \ddots & \vdots \\ H_{k1} & \cdots & H_{kk} \end{pmatrix}\right].$$

Định lý 5.2 (MLE hội tụ). *Gọi $\theta^* \in \Theta$ là tham số* thật *cần tìm. Giả sử*

1. *Các tham số là indentifiable*
2. *Support của $\mathbb{P}_\theta$ không phụ thuộc vào θ với mọi $\theta \in \Theta$*
3. *θ^* không nằm trên biên giới của Θ*
4. *Thông tin Fisher khả nghịch lân cận θ^**
5. *Một số điều kiện kỹ thuật khác*

Khi đó, chuỗi $\hat{\theta}_n^{MLE}$ thỏa mãn:

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}_{\theta^*}} \theta^*$$

$$\sqrt{nI(\theta^*)}\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)\ w.r.t. \mathbb{P}_{\theta^*}} \mathcal{N}(0, 1)\,.$$

Trường hợp mô hình đa biến, $\theta \in \mathbb{R}^k$:

$$n\left(\hat{\theta}_n^{MLE} - \theta^*\right)^T I(\theta^*)\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)\ w.r.t. \mathbb{P}_{\theta^*}} \chi_k^2.$$

Chú ý là điều kiện số 2 dễ bị vi phạm, ví dụ $X_i \overset{iid}{\sim} Unif[0, \theta]$ mà lại cần tìm tham số θ.

5.4 Mix model

Định nghĩa 5.7. Cho các mô hình gốc $X^{(k)}, k = 1, \dots, K$, lấy biến tiềm ẩn Z trên $\{1, \dots, K\}$ làm trọng số, ta có mô hình hỗn hợp

$$X = \sum_{k=1}^{K} \mathbb{P}(Z = k) X^{(k)}.$$

5.4.1 Giải thuật EM

Định nghĩa 5.8. (**E**stimation **M**aximization) *có thể* tìm được tham số θ của mô hình hỗn hợp ĐN 5.7.

Giả sử ta quan sát được $X_1 = x_1, \dots, X_n = x_n$. Gọi các trọng số tiềm ẩn tương ứng là $Z_1 = z_1, \dots, Z_n = z_n$.

Sau khi khởi tạo $\theta = \theta_0$ ngẫu nhiên, ta lặp lại 2 bước **E, M** như sau để cập nhật $\theta_k, k = 1, 2, \dots$ cho đến khi hội tụ.

- **E**stimate: Ước lượng $Z_i \approx \omega_i \coloneqq \mathbb{E}\left[Z | X_i = x_i, \theta = \theta_{k-1}\right], i = 1, \dots, n.$
- **M**aximize: Thay Z_i bởi ω_i vào công thức likelihood để tối ưu MLE $\theta = \theta_k$

5.5 M-estimation

Định nghĩa 5.9 (M-estimatior)**.** Để ước lượng thuộc tính μ^* của $X \sim \mathbb{P}$, ta tìm một "hàm tổn thất" $\rho(X, \mu)$ có kỳ vọng cực tiểu tại μ^* :

$$\begin{aligned} \mathcal{Q}(\mu) &\coloneqq \mathbb{E}\left[\rho(X, \mu)\right], \\ \mu^* &\equiv \underset{\mu}{\operatorname{argmin}}\, \mathcal{Q}(\mu). \end{aligned}$$

Quan sát $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}$, ta ước lượng

$$\mathcal{Q}(\mu) \approx \mathcal{Q}_n(\mu) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \mu).$$

Khi đó $\mu^* \approx \hat{\mu}$ với "M-estimator" $\hat{\mu}$ là

$$\hat{\mu} := \underset{\mu}{\operatorname{argmin}}\, \mathcal{Q}_n(\mu).$$

Ví dụ 5.1. Với $\rho(x, \theta) = -\ln p_\theta(x)$ ta có MLE để ước lượng tham số θ^* của mô hình $\mathbb{P}$.

Ví dụ 5.2. Với $\boldsymbol{x}, \boldsymbol{\mu} \in \mathbb{R}^d$, dùng $\rho(\boldsymbol{x}, \boldsymbol{\mu}) = \|\boldsymbol{x} - \boldsymbol{\mu}\|^2$ ta ước lượng được mean $\boldsymbol{\mu}^* = \mathbb{E}\left[\boldsymbol{x}\right]$.

Ví dụ 5.3. Có thể dùng M-estimation để ước lượng median và quantiles.

5.6 Method of moments

Định nghĩa 5.10 (Method of moments estimator)**.** Để ước lượng $\theta \in \Theta \subset \mathbb{R}^d$

1. Từ quan sát $X_1, \ldots, X_n$ tính tích suất mẫu (ĐN 1.19) $\hat{m}_1, \ldots, \hat{m}_d$
2. Với m_i là tích suất bậc i (ĐN 1.17), tìm $\hat{\theta} \in \mathbb{R}^d$:

$$\begin{aligned} m_1(\hat{\theta}) &= \hat{m}_1, \\ m_2(\hat{\theta}) &= \hat{m}_2, \\ &\vdots \\ m_d(\hat{\theta}) &= \hat{m}_d. \end{aligned}$$

5.7 Biến đổi tham số

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$, giả sử MLE, method of moments hoặc M-estimation ước lượng được duy nhất một giá trị tham số tối ưu là $\theta^* \in \Theta$. Áp dụng biến đổi tham số $\eta = \phi(\theta)$ (ĐL 1.2) lên các mô hình đó sẽ dẫn tới nghiệm tối ưu η^* thỏa mãn $\eta^* \equiv \phi(\theta^*)$.

6 Kiểm định

Hypothesis testing

- Định nghĩa giả thuyết không và giả thuyết đối.
- Thiết kế kiểm định thống kê
- Phân loại lỗi Loại 1 và Loại 2
- Tính hàm công suất của kiểm định
- Định mức đối với kiểm định
- Tính toán và giải thích giá trị p

6.1 Giả thuyết

Với mô hình thống kê $(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$, sử dụng bộ mẫu dữ liệu iid $X_1, \ldots, X_n$, ta xem xét hai giả thuyết về tham số θ như sau:

$$\begin{cases} H_0 : & \theta \in \Theta_0 \\ H_1 : & \theta \in \Theta_1 \end{cases}$$

với Θ_0, Θ_1 là phân mảnh (không giao nhau) của Θ, Θ_0 là "thường thức" (status quo), còn Θ_1 là "phát hiện" (discovery) mới. Ta gọi H_0 là giả thuyết không, còn H_1 là giả thuyết đối (thay thế).

6.1.1 Kiểm định

Ta sẽ **kiểm định** H_0 đối với H_1 bằng cách chọn và sử dụng một định lượng thống kê $\psi(X_1, \ldots, X_n) \in \{0, 1\}$.

	$\psi = 0$: chấp nhận H_0	$\psi = 1$: phủ nhận H_0
$\theta \in \Theta_0$	Kiểm định đúng	Lỗi loại 1
$\theta \in \Theta_1$	Lỗi loại 2	Kiểm định đúng

Có thể viết $\psi = \mathbb{1}\{R_\psi\}$ với sự kiện R_ψ là **vùng phủ nhận**, còn R_ψ^c là **vùng chấp nhận** .

Ta thiết kế kiểm định sao cho hàm **công suất** sau đây có giá trị nhỏ khi $\theta \in \Theta_0$ và lớn khi $\theta \in \Theta_1$:

$$\beta_\psi(\theta) := \mathbb{P}_\theta(\psi = 1) \equiv \mathbb{P}_\theta(R_\psi) \in [0, 1]$$

Vùng phủ nhận thường có dạng

$$R_\psi = \{X_i : T(X_i) \geq c\}$$

với T là một lượng thống kê còn c là một giá trị biên.

6.1.2 Mức độ lỗi

Kiểm định ψ là ở mức (significance level) $\alpha \in (0, 1)$ nếu có xác suất lỗi loại 1 không vượt quá α :

$$\sup_{\theta \in \Theta_0} \beta_\psi(\theta) \leq \alpha.$$

Chuỗi kiểm định $(\psi_n)_{n=1,2,...}$ được gọi là tiệm cận về mức α nếu

$$\lim_{n \to \infty} \sup_{\theta \in \Theta_0} \beta_{\psi_n}(\theta) \leq \alpha.$$

Phương thức Neyman-Pearson chọn một mức α, đảm bảo xác suất lỗi loại 1 không vượt quá α rồi tối thiểu hóa xác suất lỗi loại 2. Nói cách khác là giữ cho công suất $\beta_\psi(\theta)$ đủ nhỏ khi $\psi \in \Theta_0$, rồi tối đại hóa công suất khi $\psi \in \Theta_1$.

6.1.3 p-value

Từ quan sát $X_1, \dots, X_n$ ta tính giá trị mức α (tiệm cận) nhỏ nhất tại đó kiểm định ψ phủ nhận H_0, gọi nó là p-value (tiệm cận) của ψ. Nếu p-value càng nhỏ thì ta càng tự tin phủ nhận H_0.

$$\text{p-value} := \inf_{X_1,\dots,X_n;\theta \in H_0} \beta_\psi(\theta)$$

p-value	chứng cứ phủ nhận H_0
$< 0.1\%$	vô cùng mạnh
0.1%—1%	rất mạnh
1%—5%	mạnh
5%—10%	yếu

p-value	chứng cứ phủ nhận H_0
$> 10\%$	không có

6.1.4 Khoảng tin cậy

Thông thường ta có thể xây dựng được kiểm định từ khoảng tin cậy. Ví dụ, kiếm định tham số θ, $H_0 : \theta = \theta_0$, đối $H_1 : \theta \neq \theta_0$. Giả sử ta có khoảng tin cậy $\mathscr{I}$ ở mức $1 - \alpha$, tức là

$$\mathbb{P}_\theta(I \ni \theta) \geq 1 - \alpha.$$

Khi đó, $\psi = \mathbb{1}\{\theta_0 \notin \mathscr{I}\}$ là kiểm định mức α

$$\beta_\psi(\theta_0) = \mathbb{P}_{\theta_0}(\theta_0 \notin I) \leq \alpha.$$

6.2 Kiểm định tham số

6.2.1 Wald Test

Giả sử $\hat{\theta}$ là ước lượng của tham số θ, và $\hat{V}[\hat{\theta}]$ là ước lượng phương sai của $\hat{\theta}$, sao cho

$$\frac{\hat{\theta} - \theta}{\sqrt{\hat{V}[\hat{\theta}]}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, 1)\,.$$

Đặt

$$W := \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{V}[\hat{\theta}]}},$$

ta có thể xây dựng các kiểm định Wald có mức tiệm cận là α, tức là $\mathbb{P}_{H_0}(\psi_\alpha) \xrightarrow[n\to\infty]{} \alpha$.

Giả thuyết	Kiểm định Wald	asymp. p-value
$H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$	$\psi_\alpha = \mathbb{1}\{\lvert W\rvert > q_{\alpha/2}\}$	$\mathbb{P}(\lvert Z\rvert > \lvert W^{obs}\rvert)$
$H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$	$\psi_\alpha = \mathbb{1}\{W > q_\alpha\}$	$\mathbb{P}(Z > W^{obs})$
$H_0 : \theta \geq \theta_0, H_1 : \theta < \theta_0$	$\psi_\alpha = \mathbb{1}\{W < -q_\alpha\}$	$\mathbb{P}(Z < W^{obs})$

Trong bảng trên, p-value được tính từ W^{obs} là một quan sát đối với W.

6.2.2 Định lý Cochran

Định lý 6.1. *Giả sử $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ có phương sai mẫu không lệch (ĐN 1.16) là S_n^2. Khi đó*

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

và $\bar{X}_n, S_n^2$ độc lập với nhau.

6.2.3 Student's t-test

Giả sử $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ theo phân phối chuẩn (ĐN 2.7), μ, σ chưa biết, và cần kiểm định μ. Theo ĐL 6.1,

$$T := \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \equiv \frac{\sqrt{n}(\bar{X}_n - \mu)/\sigma}{\sqrt{S_n^2/\sigma^2}} \sim t_{n-1}$$

tuân theo phân phối Student's T. Ta có thể xây dựng các kiểm định Student có mức α, tức là $\mathbb{P}_{H_0}(\psi_\alpha) \equiv \alpha$.

Giả thuyết	Kiểm định Student	p-value
$H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$	$\psi_\alpha = \mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_{n-1}}\}$	$\mathbb{P}_{t_{n-1}}(\lvert Z\rvert > \lvert T^{obs}\rvert)$
$H_0 : \mu \leq \mu_0, H_1 : \mu > \mu_0$	$\psi_\alpha = \mathbb{1}\{T > q_\alpha^{t_{n-1}}\}$	$\mathbb{P}_{t_{n-1}}(Z > T^{obs})$
$H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$	$\psi_\alpha = \mathbb{1}\{T < -q_\alpha^{t_{n-1}}\}$	$\mathbb{P}_{t_{n-1}}(Z < T^{obs})$

Trong bảng trên, p-value được tính từ T^{obs} là một quan sát đối với T.

6.2.4 Two-sample t-test

Giả sử $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_y, \sigma_y^2)$, với $\mu_x, \sigma_x, \mu_y, \sigma_y$ chưa biết, và cần kiểm định $\mu_x - \mu_y$. Đặt

$$\bar{X}_n := \frac{\sum_{i=1}^n X_i}{n}, \quad \widehat{\sigma_x^2} := \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1},$$
$$\bar{Y}_m := \frac{\sum_{i=1}^m Y_i}{m}, \quad \widehat{\sigma_y^2} := \frac{\sum_{i=1}^m (Y_i - \bar{Y}_m)^2}{m-1}.$$

Ta có Student's T (ĐN 2.10) với độ tự do N tuân theo công thức Welch-Satterthwaite:

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_x - \mu_y)}{\sqrt{\widehat{\sigma_x^2}/n + \widehat{\sigma_y^2}/m}} \sim t_N,$$

$$\begin{aligned} N &= \frac{(\widehat{\sigma_x^2}/n + \widehat{\sigma_y^2}/m)^2}{\hat{\sigma}_x^4/\big(n^2(n-1)\big) + \hat{\sigma}_y^4/\big(m^2(m-1)\big)} \\ &\geq \min(n, m) \end{aligned}$$

Đặt t-statistic

$$T_{n,m} := \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\widehat{\sigma_x^2}/n + \widehat{\sigma_y^2}/m}}.$$

Giả thuyết	Kiểm định 2 mẫu	p-value
$H_0 : \mu_x = \mu_y, H_1 : \mu_x \neq \mu_y$	$\psi_\alpha = \mathbb{1}\{\lvert T\rvert > q_{\alpha/2}^{t_N}\}$	$\mathbb{P}_{t_N}(\lvert Z\rvert > \lvert T^{obs}\rvert)$
$H_0 : \mu_x \leq \mu_y, H_1 : \mu_x > \mu_y$	$\psi_\alpha = \mathbb{1}\{T > q_\alpha^{t_N}\}$	$\mathbb{P}_{t_N}(Z > T^{obs})$
$H_0 : \mu_x \geq \mu_y, H_1 : \mu_x < \mu_y$	$\psi_\alpha = \mathbb{1}\{T < -q_\alpha^{t_N}\}$	$\mathbb{P}_{t_N}(Z < T^{obs})$

6.2.5 Kiểm định tỷ lệ hợp lý

Likelihood ratio test

Với dữ liệu $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{P}_\theta$, để kiểm định $H_0 : \theta \in \Theta_0$, $H_1 : \theta \notin \Theta_0$, định lượng tỷ lệ hợp lý là

$$T_n = 2 \ln \frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_0)}$$

với $L(\hat{\theta})$ là hợp lý cực đại tổng quát, còn $L(\hat{\theta}_0)$ là hợp lý cực đại khi H_0 đúng.

6.2.6 Định lý Wilks

Giả sử $\theta = \big(\theta_1, \dots, \theta_{q+r}\big) \in \Theta \subset \mathbb{R}^{q+r}$,

$$\Theta_0 = \Big\{\theta \in \Theta : \big(\theta_{q+1}, \dots, \theta_{q+r}\big) = \big(\theta_{q+1}^{(0)}, \dots, \theta_{q+r}^{(0)}\big)\Big\}$$

với $\big(\theta_{q+1}^{(0)}, \dots, \theta_r^{(0)}\big) \in \mathbb{R}^r$ là cố định. Nếu H_0 đúng và các điều kiện hội tụ của MLE (ĐL 5.2) được thỏa mãn thì:

$$T_n \xrightarrow[n\to\infty]{(d)} \chi_r^2.$$

6.2.7 Kiểm định nhiều lần

Gọi số lần thực hiện và quan sát kết quả kiểm định là t, mức độ lỗi loại 1 cho phép là α. Nếu H_0 là đúng thì lỗi loại 1 sẽ xuất hiện khoảng αt lần, số lần này sẽ càng lớn nếu t càng lớn.

Gọi tỷ lệ p-value không vượt quá ngưỡng α là $F(\alpha) := \mathbb{P}(\text{p-value} \leq \alpha)$. Ta có $F(\alpha) \leq \alpha$ với mọi loại kiểm định. Hơn nữa, $F(\alpha) = \alpha$ với kiểm định Student's T.

6.3 Kiểm định mô hình

6.3.1 Kiểm định χ^2

Định lý 6.2 (Kiểm định χ^2). *Cho mô hình phân loại*

$$\left(\{a_1, \dots, a_K\}, \{\mathbb{P}_{\boldsymbol{p}}\}_{\boldsymbol{p} \in \Delta_K}\right),$$

với $\Delta_K := \left\{\boldsymbol{p} \in \mathbb{R}_+^K \,:\, \sum_{i=1}^K p_i = 1\right\}$. *Đặt* $N_k = \sum_{i=1}^n \mathbb{1}\{X_i = a_k\}$. *Cố định* $\boldsymbol{p}^0 \in \Delta_K$, *kiểm định* $H_0 \,:\, \boldsymbol{p} = \boldsymbol{p}^0$ *vs* $H_1 \,:\, \boldsymbol{p} \neq \boldsymbol{p}^0$. *Nếu* H_0 *là đúng thì*

$$T_n := n \sum_{i=1}^K \frac{\left(\frac{N_i}{n} - p_i^0\right)^2}{p_i^0} \xrightarrow[n \to \infty]{(d)} \chi^2_{K-1}.$$

Định lý 6.3 (Kiểm định χ^2 cho hệ phân phối). *Xem xét hệ phân phối rời rạc* $\mathbb{P}_\theta$ *có pmf* f_θ *với* $\theta \in \Theta \subset \mathbb{R}^d$. *Với dữ liệu* $X_1, \dots, X_n$, *gọi MLE của* θ *là* $\hat{\theta}$, *ước lượng* $\hat{p}_i := f_{\hat{\theta}}(a_i)$, $i = 1, \dots, K$. *Kiểm định* $H_0 \,:\, \theta \in \Theta$ *vs* $H_1 \,:\, \theta \notin \Theta$. *Nếu* H_0 *là đúng thì*

$$T_n := n \sum_{i=1}^K \frac{\left(\frac{N_i}{n} - \hat{p}_i\right)^2}{\hat{p}_i} \xrightarrow[n \to \infty]{(d)} \chi^2_{K-d-1}.$$

T_n *không phải pivotal.*

Hệ quả 6.1 (Kiểm định χ^2 cho phân phối liên tục). *Đối với phân phối liên tục, ta có thể phân hoạch tập giá trị biến thành các khoảng (quantile) để chuyển thành phân phối rời rạc rồi áp dụng kiểm định* χ^2.

6.3.2 Kiểm định KS

Định nghĩa 6.1 (Kiểm định Kolmogorov-Smirnov)**.** Gọi F_n là cdf thực nghiệm (ĐN 4.4), F là cdf thật (ĐN 1.2) của $X_1, \ldots, X_n \overset{iid}{\sim} X$. Cho cdf liên tục F^0, kiểm định giả thuyết $H_0 : F = F^0$ v.s. $H_1 : F \neq F^0$. Định lượng kiểm định

$$T_n \coloneqq \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F^0(t) \right|$$

là pivotal.

Theo định lý Donsker (ĐL 4.2), nếu H_0 là đúng thì $T_n \xrightarrow[n \to \infty]{(d)} Z$, với Z là phân bố của đỉnh tuyệt đối của Brownian bridge.

Vì các cdf đều tăng đơn điệu, cdf thực nghiệm có dạng bậc thang, ta có thể xếp $X_1 \leq \ldots \leq X_n$ từ nhỏ đến lớn để tính T_n dễ dàng hơn.

$$T_n \equiv \sqrt{n} \max_{i=1,\ldots,n} \left\{ \max \left\{ \left| \frac{i-1}{n} - F^0(X_i) \right|, \left| \frac{i}{n} - F^0(X_i) \right| \right\} \right\}.$$

6.3.3 Kiểm định KL

Định nghĩa 6.2 (Kolmogorov-Lilliefors test)**.** Trong tiền đề của ĐN 6.1, nếu bỏ điều kiện F_0 cố định, thay vào đó là kiểm chứng phân phối $\mathbb{P}$ thuộc hệ phân phối chuẩn $\{\mathcal{N}\left(\mu, \sigma^2\right)$. Gọi MLE là $\hat{\mu}, \hat{\sigma}^2$, định lượng kiểm định

$$T_n \coloneqq \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - \Phi_{\hat{\mu}, \hat{\sigma}^2}(t) \right|$$

là pivotal.

7 Bayesian

7.1 Phương pháp

Định nghĩa 7.1 (Bayesian method)**.** Với dữ liệu $X_1, \dots, X_n$, ta ước lượng tham số θ cho mô hình xác suất qua các bước:

1. Gán xác suất tiên nghiệm (prior) $\pi(\theta)$ cho tham số $\theta \in \Theta$.
2. Chọn xác suất $f(X|\theta)$ phụ thuộc vào θ, gọi MLE (ĐN 5.5) là L_n
3. Tính xác suất hậu nghiệm (posterior) theo Bayes ĐL 1.3:

$$\pi(\theta|X_1, \dots, X_n) \propto L_n(X_1, \dots, X_n|\theta)\pi(\theta), \quad \theta \in \Theta.$$

7.2 Loại prior

Định nghĩa 7.2 (Conjugate prior)**.** Nếu xác suất tiền nghiệm và xác suất hậu nghiệm thuộc cùng một hệ xác suất, ta nói xác suất tiền nghiệm là liên hợp đối với mô hình.

Ví dụ 7.1. Beta là xác suất tiền nghiệm liên hợp đối với mô hình Bernoulli Ber, nhị thức Binom và hình học Geom.

Định nghĩa 7.3 (Non informative prior)**.** Nếu ta không có thông tin tiền nghiệm gì về tham số $\theta \in \Theta = [a, b]$ thì có thể dùng phân phối đồng đều $\mathrm{Unif}[a, b]$ làm xác suất tiền nghiệm.

Định nghĩa 7.4 (Improper prior)**.** Nếu hàm $\pi : \Theta \to [0, +\infty)$ là đo được (measurable) nhưng không khả tích toàn phần trên Θ thì ta nói π là improper prior.

7.3 Jeffreys prior

Định nghĩa 7.5 (Jeffreys prior)**.** Với mô hình thống kê dựa trên tham số θ cho X có thông tin Fisher (ĐN 5.6) là $I(\theta)$ thì Jeffreys prior

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)},$$

là non informative.

Ví dụ 7.2.

- Thí nghiệm $\mathrm{Ber}(p)$, $p \in (0,1)$ có

$$\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}} \propto \mathrm{B}\frac{1}{2}\frac{1}{2}.$$

- Thí nghiệm $\mathcal{N}\left(\mu, \sigma^2\right), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$ có $\pi_J(\mu, \sigma^2) \propto \sigma^{-3}$ là improper prior.

Định lý 7.1 (Jeffreys prior là invariant khi đổi tham số)**.** *Nếu đổi tham số $\theta \mapsto \eta = \phi(\theta)$ (ĐL 1.2) thì mật độ của tham số mới η là*

$$\tilde{\pi}(\eta) = \frac{\pi(\phi^{-1}(\eta))}{\phi'(\phi^{-1}(\eta))} \propto \sqrt{\det \tilde{I}(\eta)}$$

với $\tilde{I}(\eta)$ là thông tin Fisher của mô hình dựa trên tham số η thay vì θ.

Remark. Phân phối tích lũy

$$F_{\tilde{\pi}}(\eta) \equiv F_{\pi}(\phi^{-1}(\eta)).$$

7.4 Bayesian estimation

Định nghĩa 7.6 (Posterior mean)**.** Từ tiền nghiệm $\pi(\theta)$, sau khi tính hậu nghiệm $\pi(\theta|X_1, \ldots, X_n)$ ta ước lượng θ bằng trung bình:

$$\hat{\theta}^{(\pi)} := \int_{\Theta} \theta \pi(\theta|X_1, \ldots, X_n) d\theta.$$

Định nghĩa 7.7 (MAP, maximum a posteriori)**.** Chọn cực đại

$$\hat{\theta}^{\mathrm{MAP}} := \underset{\theta \in \Theta}{\operatorname{argmax}}\, \pi(\theta|X_1, \ldots, X_n).$$

Định lý 7.2 (Asymptotic properties of the Bayes estimator)**.** *Thông thưởng, tính hội tụ về chuẩn (ĐN 3.4) của Bayes estimator không phụ thuộc vào chọn lựa xác suất tiền nghiệm.*

Remark. Tính phương sai của hội tụ tương tự như MLE (ĐL 5.2).

8 Hồi quy

Giả sử $(X_i, Y_i)_{i=1,\ldots,n} \overset{iid}{\sim} \mathbb{P}_{(X,Y)}$ với mật độ xác suất $h(x, y)$, mật độ xác suất biên $h(x)$ và mật độ xác suất có điều kiện $h(y|x)$.

Định nghĩa 8.1 (Regression function)**.** Hồi quy là kỳ vọng về Y khi biết X :

$$x \mapsto f(x) := \mathbb{E}\left[Y|X = x\right] = \int yh(y|x)dy.$$

Định nghĩa 8.2 (Conditional quantile)**.** Cho $\alpha \in [0, 1]$, phân vị q_α cho Y khi biết $X = x$:

$$x \mapsto q_\alpha(x) \text{ such that } \int_{-\infty}^{q_\alpha(x)} h(y|x)dy \equiv 1 - \alpha.$$

Đặt $\alpha = 1/2$ ta có conditional median.

8.1 Hồi quy tuyến tính

Định nghĩa 8.3 (Theoretical linear regression)**.** Giả sử $\mathrm{Var}\left(X\right) > 0$. Đường hồi quy tuyến tính áp Y lên X là $y = ax + b$ với hệ số

$$(a, b) := \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \mathbb{E}\left[(Y - \alpha X - \beta)^2\right].$$

Remark. Lấy đạo hàm từng phần và giải hệ phương trình ta có

$$a = \frac{\mathrm{Cov}\left(X, Y\right)}{\mathrm{Var}\left(X\right)},$$
$$b = \mathbb{E}\left[Y\right] - a\mathbb{E}\left[X\right].$$

Định nghĩa 8.4 (Noise)**.** Nhiễu

$$\epsilon := Y - (aX + b),$$

có $\mathbb{E}[\epsilon] \equiv 0, \operatorname{Cov}(X, \epsilon) \equiv 0$.

8.2 Bình phương tối thiểu

Định nghĩa 8.5 (Least squares estimator, LSE)**.** Quan sát $(X_i, Y_i), i = 1, \ldots, n$. LSE là điểm cực tiểu của tổng bình phương lỗi (sum of squared errors):

$$(\hat{a}, \hat{b}) := \underset{(\alpha,\beta)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \alpha X_i - \beta)^2.$$

Remark. Lấy đạo hàm từng phần và giải hệ phương trình ta có

$$\begin{aligned} \hat{a} &= \frac{\overline{XY} - \bar{X}\bar{Y}}{\overline{X^2} - \bar{X}^2}, \\ \hat{b} &= \bar{Y} - \hat{a}\bar{X}. \end{aligned}$$

Tham khảo

Giáo trình

Philippe Rigollet, Tyler Maunu, Jan Christian Huetter. 2022. "Fundamentals of Statistics". MITx. https://www.edx.org/course/fundamentals-of-statistics.
Pishro-Nik, H. 2014. *Introduction to probability, statistics, and random processes*. Kappa Research LLC. https://www.probabilitycourse.com.
Wasserman, Larry. 2004. *All of statistics : a concise course in statistical inference*. New York: Springer. https://archive.org/details/springer_10.1007-978-0-387-21736-9.

MITx 18.6501x

"Fundamentals of Statistics" (MITx 18.6501x) là khóa học của Philippe Rigollet (2022) đại học MIT dạy trên edX.

Quy định

Kỳ hạn

- Exercises and homework: Wednesdays 11:59AM UTC (Wed. 20:59 JST)
- Exams (48 hours): Tuesdays 11:59AM UTC(Tue. 20:59 JST)

Thời gian cần thiết

Mỗi tuần hơn 12 tiếng

- 5-7 hours on exercises, including 3 hours of lecture clips
- 1-2 hours watching recitations
- 5-7 hours for weekly problem sets

Tính điểm

Điểm thi đậu chứng chỉ là 60% tổng số điểm tối đa.