HỌC VIỆN KỸ THUẬT QUÂN SỰ
KHOA CÔNG NGHỆ THÔNG TIN

PGS TS TÔ VĂN BAN
ThS Phan Thu Hà

ĐỀ CƯƠNG CHI TIẾT BÀI GIẢNG

XÁC SUẤT THỐNG KÊ

(*Dùng cho hệ Dài hạn 5 năm*)

Hà nội, 9-2014

BỘ MÔN DUYỆT **Chủ nhiệm Bộ môn** Tô Văn Ban	**ĐỀ CƯƠNG CHI TIẾT BÀI GIẢNG** (Dùng cho hệ dài hạn, 60 tiết giảng) Học phần: XÁC SUẤT THỐNG KÊ Nhóm môn học: Toán Ứng dụng Bộ môn: Toán Khoa: Công nghệ Thông tin	**Thay mặt nhóm môn học** Phan Thu Hà

Thông tin về giáo viên

TT	Họ tên giáo viên	Học hàm	Học vị
1	Tô Văn Ban	Phó giáo sư	TS
3	Phan Thu Hà	Giảng viên	ThS

Địa điểm làm việc: Bộ Môn Toán, P1301, Nhà S4, 236 Hoàng Quốc Việt
Điện thoại, email: 069 515 330, bomontoan_hvktqs@yahoo.com

Bài giảng 1: Biến cố và xác suất của biến cố

Chương, mục: 1

Tiết thứ: 1- 4 Tuần thứ: 1

Mục đích, yêu cầu:

- Nắm sơ lược về Học phần, các quy định chung, các chính sách của giáo viên, các địa chỉ và thông tin cần thiết, bầu lớp trưởng Học phần.
- Nắm được, tính được các xác suất ở những mô hình đơn giản. Đặc biệt, vận dụng công thức xác suất toàn phần, công thức Bayes, công thức Bernoulli.
- Thấy được tính độc lập của các biến cố là đặc thù của lý thuyết XS

- Hình thức tổ chức dạy học:
Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:
Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:
Giảng đường do P2 phân công.

- Nội dung chính:

Giới thiệu về môn học và các quy định

Chương 1. Biến cố và xác suất của biến cố

§1.1.Xác suất biến biến cố

§1.2. Xác suất điều kiện

§1.3. Sự độc lập

Giới thiệu học phần XSTK(15 phút)

- Xuất phát điểm của Lý thuyết xác suất là tung đồng tiền, đánh bạc hay các trò chơi may rủi.

- Nhiều nghịch lý được phát hiện dẫn đến những tranh cãi kịch liệt ở thế kỷ 19, dẫn đến luồng quan điểm coi lý thuyết xác suất là "khoa học ngây thơ".

- Do nhu cầu phát triển như vuc bão của khoa học ở đầu thế kỷ 20, do đòi hỏi của vật lý, thiên văn, sinh học…, dựa trên lý thuyết tập hợp và lý thuyết độ đo đã rất phát triển, Kolmogrov, nhà bác học Nga hàng đầu đã đưa ra hệ tiên đề của LTXS, làm cơ sở toán học vững chắc cho ngành toán học này. Lý thuyết XS là cơ sở của thống kê toán, một ngành của toán học được ứng dụng rộng rãi nhất hiện nay.

- Trên thế giới, thống kê rất được phát triển. Nhiều khoa toán nằm trong trường thống kê.

- Chia làm 2 phần: Phần XS gồm 3 chương, phần thống kê gốm 2 chương

Chính sách riêng

Mỗi lần lên bảng chữa bài tập đúng được ghi nhận, cộng vào điểm quá trình 0.5 điểm. Chữa bài tập sai không bị trừ điểm.

Sự hiện diện trên lớp: Không đi học ≥ 5 buổi sẽ không được thi.

Tài liệu tham khảo cho Học phần LTXSTK

TT	Tên tài liệu	Tác giả	Nxb	Năm xb
1	Xác suất thống kê, Tô Văn Ban,	Tô Văn Ban	Nxb Giáo dục Việt Nam	2010
2	Xác suất Thống kê	Tống Đình Quỳ	Giáo dục	2006
3	Mở đầu về lý thuyết Xác suất và các ứng dụng	Đặng Hùng Thắng	Giáo dục	2005
4	Lý thuyết Xác suất	Nguyễn Xuân Viên	HV KTQS	1998
5	Thống kê và ứng dụng	Đặng Hùng Thắng	Giáo dục	1999

Đề Bài tập về nhà XSTK

(Gạch dưới: Chữa trên lớp)

CHƯƠNG I
Tài liệu [1]: 1(2 – 3 – 5 – 7 – 9 – 10 - 11–13 – 15 – 17 – 18 –19 - 20 – 21 – 22 – 23 – 24 - 27 -29).
Tài liệu [2]: Tr 35-38: 6, 9, 10, 12, 13, 15, 21, 25, 29, 30, 33 (sửa 10% thành 7%).

CHƯƠNG II
Tài liệu [1]: 2(1 - 2 –3 - 4– 5- 6 - 7 - 8 – 9 - 10 - 11 –12- 14 – 16-17 - 18-21-23- 26 - 27- 30-32). Tài liệu [2]: Tr 76-78: 2, 4, 8 (sửa x thành |x|), 10.
CHƯƠNG III
Tài liệu [1]: 3(1 – 3 – 4 – 6 – 8 – 9 -10- 11 – 21 – 22 – 24 -26- 27 – 33 – 38- 40-49- 53 - 54- 55). Tài liệu [2]: Tr 110-112: 10, 11, 14, 15, 16.
CHƯƠNG IV
Tài liệu [1]: 4(1 – 4 – 5 – 6 – 10 – 11 – 12 – 13 – 14 – 17 – 19 – 21 – 23 – 24 – 25(a) – 26(a,b) – 27 – 29 – 30 – 31 – 32 –33- 34 – 35 – 37). Tài liệu [2]: Tr 153-157: 11, 12,15,17, 19,22.
CHƯƠNG V Tài liệu [1]: 5(1- 4 - 5- 6- 8 - 9- 12- 14 - 15) Tài liệu [2]: Tr187-189: 3, 4, 6, 8, 10, 14, 16, 17, 22, 26, 28

CẤU TRÚC ĐỀ THI, CÁCH THỨC CHO ĐIỂM

Câu số	Về phần	Số điểm
1	Lý thuyết	2.0
2	Chương 1	2.0
3	Chương 2, chương 3	2.0
4	Chương 4	2.0
5	Kiểm định độc lập, TQ, HQ	2.0
Điểm bài thi		10đ
Điểm quá trình		10đ
Điểm chuyên cần		10đ
Tổng điểm = điểm chuyên cần x 10% + điểm quá trình x 20% + điểm bài thi x 70%		10đ

Hình thức thi: Thi viết

Bầu lớp trưởng lớp học phần. Kết quả:

Số điện thoại giáo viên:

Địa chỉ Email cần:

Webside cần:

Chương 1
BIẾN CỐ, XÁC SUẤT BIẾN CỐ

§ 1.1. XÁC SUẤT BIẾN CỐ (2 tiết)

1.1.1.Thí nghiệm ngẫu nhiên, biến cố, không gian mẫu

Định nghĩa. Thí nghiệm ngẫu nhiên là thí nghiệm ở đó kết quả ở đầu ra không được xác định duy nhất từ những hiểu biết về đầu vào.

Kết quả ở đầu ra của thí nghiệm được quy định là kết quả đơn, không phân tách được, mỗi lần thử chỉ có một kết quả. Vì thế ta hay gọi chúng là những kết cục (hay biến cố sơ cấp), ký hiệu bởi ζ hay thêm vào chỉ số: $\zeta_1, \zeta_2, \ldots$

Tập tất cả những kết cục có thể có của một thí nghiệm ngẫu nhiên, ký hiệu bởi S (nhiều tài liệu viết là Ω), được gọi là không gian mẫu (hay tập vũ trụ) của thí nghiệm đó.

Hợp thành của các kết cục nào đó, chính là 1 tập con của S, được gọi là một biến cố. Bản thân tập S cũng là một biến cố, được gọi là biến cố chắc chắn. Biến cố trống không chứa bất cứ kết cục nào, ký hiệu bởi $\varnothing$, được gọi là biến cố bất khả (hay biến cố không thể). Biến cố $\{\zeta\}$ gồm một kết cục ζ được gọi là biến cố sơ cấp, để đơn giản vẫn được ký kiệu là ζ. Các biến cố được ký hiệu bởi chữ cái in hoặc thêm chỉ số: $A, B, \ldots, A_1, A_2, \ldots$ Chúng ta có thể thể hiện biến cố bằng cách liệt kê các kết cục hoặc nêu các thuộc tính của nó, tất cả được viết trong dấu ngoặc nhọn { . }. Nếu kết quả của lần thử nào đó là ζ và $\zeta \in A$ thì ta nói biến cố A xảy ra ở lần thử này.

Không gian mẫu có một số hữu hạn hoặc đếm được các kết cục được gọi là không gian mẫu rời rạc; trái lại, không gian mẫu được gọi là liên tục.

Ví dụ 1.1. Tìm không gian mẫu của thí nghiệm tung đồng tiền

i)1 lần; ii) 2 lần.

Giải. i) Hai kết cục có thể: ngửa N và sấp S. Vậy $S = \{N, S\}$.

ii) $S = \{NN, NS, SN, SS\}$.

Như vậy ở trường hợp ii) không gian mẫu có 4 kết cục, cũng có đúng 4 biến cố sơ cấp. Cả thảy gồm $2^4 = 16$ biến cố:

$\varnothing, \{NN\}, \{NS\}, \{SN\}, \{SS\}, \{NN, NS\}, \ldots, \{NN, NS, SN, SS\}$.

Nói chung, nếu không gian mẫu có N kết cục thì có cả thảy 2^N biến cố.

Một số biến cố quan tâm có thể là:

A = {ngửa ở lần đầu} = { NN, NS}

B = {chỉ có 1 lần ngửa} = {NS, SN},

C = {ít nhất 1 lần ngửa} = {NN, NS, SN}, … #

Ví dụ 1.2. Tung đồng tiền đến khi xuất hiện mặt sấp thì dừng lại.

Đối với thí nghiệm này chúng ta đặt

$\zeta_1 = S, \quad \zeta_2 = NS, \ldots, \ \zeta_n = NN\ldots NS$ (n – 1 lần N).

Không gian mẫu là $S = \{\zeta_1, \zeta_2, \ldots, \zeta_n, \ldots\}$.

Tuy nhiên, nếu ta chỉ quan tâm đến số lần tung đồng tiền cần thiết thì có thể xét không gian mẫu là $S = \{1, 2, 3, \ldots\}$.

Với thí nghiệm này, không gian mẫu gián đoạn, có vô hạn kết cục. Một số biến cố quan tâm có thể là:

A = {số lần tung là chẵn}, C = {số lần tung từ 5 đến 10},

B = {số lần tung < 10}, D = {số lần tung bằng $1,4\} = \varnothing$.#

Ví dụ 1.3. Các chíp điện tử được sản xuất bằng cách cấy các ion vào sâu trong màng silicon dioxide (S_iO_2). Quá trình cấy mang bản chất ngẫu nhiên, một số ion vào sâu hơn so với dự định, số khác thì không. Thí nghiệm ngẫu nhiên có thể xét đến ở đây là độ sâu (theo μm) của ion được cấy vào màng silicon thế nào. Vậy có thể chọn $S = [0; 20]$. Không gian mẫu vô hạn, hơn nữa liên tục. #

Chúng ta muốn gán mỗi biến cố A với một số - ký hiệu là P(A), gọi là xác suất của biến cố A - đặc trưng cho khả năng xảy ra biến cố A trong mỗi lần thử. Việc gán đó phải thoả mãn các tính chất tự nhiên sau đây.

$$P(A) \geq 0. \qquad (1.1.1)$$

$$P(S) = 1. \qquad (1.1.2)$$

Nếu A và B là hai biến cố xung khắc thì

$$P(A + B) = P(A) + P(B). \qquad (1.1.3)$$

1.1.2. Định nghĩa cổ điển về xác suất

Giả sử đối với một thí nghiệm ngẫu nhiên nào đó có cả thảy N kết cục và chúng là đồng khả năng. Hơn nữa, giả sử rằng có n_A kết cục là thuận lợi cho biến cố A (nghĩa là biến cố A xảy ra khi và chỉ khi một trong các kết cục này xảy ra). Xác suất của biến cố A được xác định bởi

$$P(A) = \frac{n_A}{N} = \frac{\text{Sè kÕt côc thuËn lîi}}{\text{Tæng sè kÕt côc ®ång kh¶ n¨ng}} \qquad (1.1.4)$$

Ví dụ 1.4. Trong bình có a quả cầu trắng, b quả cầu đen $(a > 0, b > 0)$ với trọng lượng, kích thước giống hệt nhau. Lắc đều rồi lấy ngẫu nhiên 1 quả. Tìm xác suất để quả cầu lấy được có màu trắng.

Giải. Rõ ràng số kết cục đồng khả năng là a + b. Đặt A = {rút được quả cầu trắng} thì có a kết cục thuận lợi cho A (A xảy ra khi và chỉ khi rút được 1 trong a quả cầu trắng). Từ định nghĩa $P(A) = a/(a+b)$.
#

Ví dụ 1.5. Một hộp có 10 sản phẩm trong đó có 6 chính phẩm (và 4 phế phẩm). Lấy ngẫu nhiên từ hộp 3 sản phẩm. Tìm xác suất để:

i) Cả 3 sản phẩm đều chính phẩm; ii) Có đúng 2 chính phẩm.

Giải. Đặt A = {cả 3 sản phẩm rút được đều là chính phẩm};
B = {Rút được đúng 2 chính phẩm}.

Số kết cục đồng khả năng chính là số cách rút 3 sản phẩm từ 10 sản phẩm hay C_{10}^3 cách.

i) Số kết cục thuận lợi cho A là C_6^3. Vậy

$$P(A) = C_6^3 / C_{10}^3 = 1/6 \approx 0{,}167 \ (= 16{,}7\%).$$

ii) Hai chính phẩm được rút trong 6 chính phẩm, vậy có C_6^2 cách.

Một phế phẩm được rút trong 4 phế phẩm, vậy có C_4^1 cách.

Số kết cục thuận lợi cho B là $C_6^2 C_4^1$. Vậy $P(B) = \dfrac{C_6^2 C_4^1}{C_{10}^3} = \dfrac{1}{2}$.#

Ví dụ 1.6. Trong một cuộc liên hoan một tổ gồm 10 người ngồi quanh một chiếc bàn tròn một cách ngẫu ngiên. Tìm xác suất để tổ trưởng A và tổ phó B ngồi cạnh nhau.

Giải. Chúng ta đánh số ghế ngồi từ 1 đến 10 và coi 2 cách ngồi là khác nhau nếu có ít nhất 1 chỗ thấy có 2 người ngồi khác nhau.

Số kết cục (đồng khả năng) là 10! (10 người ngồi vào 10 chỗ).

Để tính số kết cục thuận lợi, ta xếp A ngồi tuỳ ý vào 1 trong 10 chỗ (10 cách); B ngồi vào 1 trong 2 chỗ cạnh A (2 cách); 8 người còn lại ngồi tuỳ ý vào 8 chỗ còn lại (8! cách). Số kết cục thuận lợi là 10. 2. 8!. Ta nhận được

$$P(B) = 10.2.8!/10! = 2/9 . \qquad \#$$

Xác suất hình học. Nếu thí nghiệm ngẫu nhiên có thể cho tương ứng với việc gieo ngẫu nhiên 1 điểm tuỳ ý trên miền hình học G sao cho khả năng để điểm đó rơi vào miền $g \subset G$ tỷ lệ với diện tích của miền này, không phụ thuộc vào vị trí tương đối của g với G cũng như vào hình dạng của nó. Khi đó, xác suất biến cố A cho bởi

$$P(A) = \frac{\text{Sè ®o miÒn } g_A}{\text{Sè ®o miÒn } G} \qquad (1.1.5)$$

trong đó g_A : miền ứng với biến cố A,

số đo: độ dài, diện tích, thể tích (tương ứng trong $\mathbb{R}^1$, $\mathbb{R}^2$, $\mathbb{R}^3$).

Nhận xét. Trong định nghĩa cổ điển các phép thử chỉ là giả định, ta không phải thực hiện bất kỳ phép thử nào; các xác suất là tiên nghiệm, được suy đoán một cách lôgíc từ tính đối xứng. Định nghĩa thoả mãn các đòi hỏi (1.1.1) - (1.1.3).

Tuy nhiên, định nghĩa có nhiều nhược điểm. Trong định nghĩa có từ đồng khả năng, một trong những khái niệm mà ta đang cần xây dựng. Như đã thấy, điều này gây khó khăn khi xác định n_A và N.

Mặc dầu đã cải thiện tình hình, song xác suất hình học vẫn chưa giải quyết được trường hợp các kết cục không đồng khả năng.

1.1.3. Định nghĩa xác suất bằng tần suất

Lặp lại một thí nghiệm nào đó n lần và giả sử biến cố A đã cho xuất hiện n_A lần. Số n_A được gọi là tần số, còn tỷ số $\frac{n_A}{n}$ được gọi là tần suất (hay tần số tương đối) xuất hiện biến cố A trong n lần thử đó.

Ví dụ 1.9. Tiến hành tung đồng tiền cân đối một cách vô tư nhất người ta thu được kết quả

Người làm thí nghiệm	Số lần tung	Số lần mặt sấp	Tần suất
Buffon	4 040	2 048	0,5069
Pearson	12 000	6 019	0,5016
Pearson	24 000	12 012	0,5005

Khi số phép thử tăng lên vô hạn, ta hy vọng tần suất dần đến 0,5, số này được lấy làm xác suất của biến cố hiện mặt sấp khi tung đồng tiền 1 lần. #

Nói chung, tần suất thay đổi từ loạt thử này sang loạt thử khác. Tuy nhiên khi n tăng, tần suất có tính ổn định, nó dường như dao động quanh số p nào đó. Số cố định p đó được xem là xác suất của biến cố A.

Định nghĩa. Giới hạn của tần suất $\frac{n_A}{n}$ khi n tăng lên vô hạn được gọi là xác suất của biến cố A theo nghĩa thống kê (hay theo tần suất):

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}. \qquad (1.1.6)$$

Theo định nghĩa này, khi n lớn, ta có thể dùng xấp xỉ

$$P(A) \approx \frac{n_A}{n}. \qquad (1.1.7)$$

1.1.4. Mối quan hệ giữa các biến cố, phép toán trên biến cố

Như đã nói, không gian mẫu S là tập tất cả các kết cục ζ của một thí nghiệm ngẫu nhiên. Mỗi tập con của S là một biến cố; bản thân S là biến cố, gọi là biến cố chắc chắn. Biến cố không thể ký hiệu là $\varnothing$.

a) Hợp các biến cố.

Biến cố C gọi là hợp của hai biến cố A và B và ta viết $C = A + B$ hoặc $C = A \cup B$, nếu trong một lần thử bất kỳ (sau đây để đơn giản ta sẽ bỏ cụm từ này), biến cố C xảy ra khi và chỉ khi hoặc A, hoặc B, hoặc cả A và B đều xảy ra (xem lược đồ Venn ở Hình 1.2(a)).

Chúng ta dễ dàng hiểu ý nghĩa hợp của n biến cố, được ký hiệu bởi một trong những cách sau:

$$A_1 + A_2 + \ldots + A_n;\ A_1 \cup A_2 \cup \ldots \cup A_n;\ \sum_{i=1}^{n} A_i;\ \bigcup_{i=1}^{n} A_i.$$

b) Kéo theo. Biến cố A được gọi là kéo theo biến cố B, ký hiệu $A \subset B$, nếu biến cố A xảy ra thì biến cố B xảy ra (xem lược đồ Venn ở Hình 1.2(b)).

c) Biến cố xung khắc. Hai biến cố A và B được gọi là xung khắc nếu biến cố A xảy ra thì biến cố B không xảy ra và ngược lại, nếu biến cố B xảy ra thì biến cố A không xảy ra (xem Hình 1.2(c)).

Tổng quát, các biến cố $A_1, A_2, ..., A_n$ được gọi là xung khắc từng đôi nếu bất kỳ 2 biến cố nào trong chúng cũng là xung khắc.

d) Biến cố đối. Biến cố B được gọi là biến cố đối (hay phần bù) của biến cố A, và ta viết $B = \overline{A}$, nếu chúng xung khắc và hợp của chúng là biến cố chắc chắn (xem Hình 1.2(d)). Như vậy,

$$A \cup \overline{A} = S;\ \ A, \overline{A}\ \text{ xung khắc. Rõ ràng, }\ \overline{\overline{A}} = A.$$

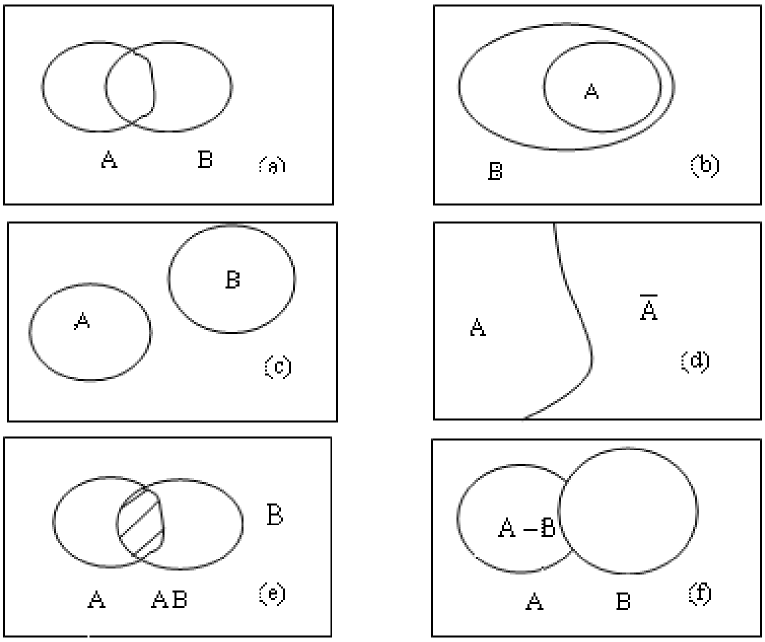
e) Giao 2 biến cố. Biến cố C gọi là giao (hay tích) của hai biến cố A và B, và ta viết $C = A \cap B$ (hay $C = AB$) nếu C xảy ra khi và chỉ khi cả A và B đều xảy ra (xem Hình 1.2(e)).

Tổng quát, biến cố A được gọi là giao (hay tích) của các biến cố $A_1, ..., A_n$ nếu A xảy ra khi và chỉ khi mọi biến cố $A_1, ..., A_n$ đều xảy ra.

Tích các biến cố được ký hiệu bởi một trong những cách sau:

$$A_1A_2...A_n;\ A_1 \cap A_2 \cap ... \cap A_n;\ \prod_{i=1}^{n} A_i;\ \bigcap_{i=1}^{n} A_i.$$

f) Hiệu 2 biến cố. $A - B$, (xem Hình 1.2(f)).



Hình 1.2. Lược đồ Venn: (a) hợp 2 biến cố; (b) kéo theo; (c) xung khắc; (d) biến cố đối; (e) giao 2 biến cố; (f) hiệu 2 biến cố.

Quy tắc Đờ Moocgăng (De Morgan):

i) $\overline{A \cup B} = \overline{A} \cap \overline{B};\quad \overline{A \cap B} = \overline{A} \cup \overline{B};$ (1.1.8)

ii) $B - \left(\bigcup_{i=1}^{n} A_i\right) = \bigcap_{i=1}^{n}(B - A_i);\ B - \left(\bigcap_{i=1}^{n} A_i\right) = \bigcup_{i=1}^{n}(B - A_i);$ (1.1.9)

iii) $B - \left(\bigcup_{i=1}^{\infty} A_i\right) = \bigcap_{i=1}^{\infty}(B - A_i);\ B - \left(\bigcap_{i=1}^{\infty} A_i\right) = \bigcup_{i=1}^{\infty}(B - A_i).$ (1.1.10)

Ví dụ 1.10. Rút 1 quân bài tú lơ khơ. Xét các biến cố:

A={Rút được quân đen}; B={rút được quân đỏ};

C={Rút được quân cơ có số}; D={ Rút được quân cơ từ 9 trở lên}.

Khi đó, A và C xung khắc; $B = \overline{A}$, $C \subset B$;

$C \cup D =$ {Rút được quân cơ}; $C \cap D =$ {Rút được 9 cơ hoặc 10 cơ};

$C - D =$ { Rút được quân cơ từ 2 đến 8}… #

1.1.5. Định nghĩa xác suất theo tiên đề

a) σ- đại số.

Định nghĩa. Họ $\mathcal{F}$ khác trống các biến cố của không gian mẫu S được gọi là một đại số (hay trường) nếu nó thoả mãn các tính chất: i) $S \in \mathcal{F}$;

ii) $A \in \mathcal{F} \Rightarrow \overline{A} \in \mathcal{F}$;

iii) $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$.

Ý tưởng của tính chất ii) và iii) là ở chỗ, mỗi đại số đóng với phép lấy phần bù, lấy hợp. Ngoài ra, bằng những suy diễn đơn giản và từ quy tắc De Morgan ta thấy rằng $\varnothing \in \mathcal{F}$. $\mathcal{F}$ cũng đóng với phép lấy hợp, giao, phần bù một số hữu hạn lần các phần tử của nó theo một thứ tự bất kỳ.

Ví dụ, nếu $A, B, C, D \in \mathcal{F}$ thì

$$\overline{(B+C)} \cup [D \cup (C-A)] - \overline{A \cap D} \in \mathcal{F}.$$

Định nghĩa. Nếu ngoài các tính chất i-iii, họ $\mathcal{F}$ còn có tính chất

$$\text{iii}')\quad A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F},$$

thì $\mathcal{F}$ được gọi là σ- đại số (hoặc σ-trường).

Ví dụ 1.11. Nhóm các biến cố $\{A_1, \ldots, A_n\}$ được gọi là đầy đủ nếu:

i) Chúng xung khắc từng đôi: $A_i A_j = \varnothing \ (i \neq j)$;

ii) Hợp của chúng là biến cố chắc chắn: $A_1 \cup \ldots \cup A_n = S$.

Nếu nhóm các biến cố $\{A_1, \ldots, A_n\}$ là đầy đủ thì đại số sinh bởi nhóm này rất đơn giản: Mỗi phần tử của đại số đó là hợp một số hữu hạn các biến cố nào đó trong họ đã cho. #

Ví dụ 1.12 *(σ- đại số Borel)* (xem [1])

b) Các tiên đề xác suất.

Định nghĩa. Giả sử (S, $\mathcal{F}$) là bộ gồm không gian mẫu S và đại số $\mathcal{F}$ các biến cố của S. Xác suất P(.) là một hàm tập trên $\mathcal{F}$ và thoả mãn các tiên đề sau đây:

I. $P(A) \geq 0, \quad \forall A \in \mathcal{F}$ (1.1.11)

II. $P(S) = 1$ (1.1.12)

III. $A, B \in \mathcal{F}$; A, B xung khắc thì

$$P(A+B) = P(A) + P(B). \qquad (1.1.13)$$

Trong trường hợp $\mathcal{F}$ là σ- đại số, thay cho III là

IIIa. Nếu dãy các biến cố $A_1, A_2, \ldots$ xung khắc từng đôi thì

$$P(A_1 + A_2 + \ldots) = P(A_1) + P(A_2) + \ldots \qquad (1.1.14)$$

Định nghĩa. Bộ ba (S, $\mathcal{F}$, P) bao gồm không gian mẫu S, đại số (hay σ- đại số) $\mathcal{F}$ và xác suất P(.) được gọi là không gian xác suất.

Mỗi thí nghiệm ngẫu nhiên được mô hình hoá bởi một không gian xác suất (S, $\mathcal{F}$, P) nào đó.

Từ nay, khi nói đến biến cố A nào đó thì ta hiểu đó là phần tử của họ các đại số hoặc σ- đại số nào đó trong không gian xác suất nào đó.

Cũng có thể thấy rằng, định nghĩa xác suất theo tần suất là trường hợp riêng của xác suất theo tiên đề.

1.1.6. Các tính chất của xác suất

1) $P(\varnothing) = 0$

2) $P(\overline{A}) = 1 - P(A)$; $P(A) = 1 - P(\overline{A})$.

3) A và B xung khắc thì

$$P(A+B)=P(A)+P(B).$$

3a) $A_1,...,A_n$ xung khắc từng đôi thì

$$P(A_1+...+A_n)=P(A_1)+...+P(A_n).$$

3b) A, B tuỳ ý thì

$$P(A+B)=P(A)+P(B)-P(AB).$$

3c) A,B, C tuỳ ý thì

$$P(A+B+C)=P(A)+P(B)+P(C)$$
$$-[P(AB)+P(AC)+P(BC)]+P(ABC).$$

3d) $\forall A_1,...,A_n$, $P\left(\sum_{i=1}^{n} A_i\right)=\sum_{i=n}^{n} P(A_i)-\sum_{1\le i<j\le n} P(A_iA_j)+$

$$+\sum_{1\le i<j<k\le n} P(A_iA_jA_k)-...+(-1)^{n-1}P(A_1...A_n).$$

4) $A\subset B \Rightarrow P(A)\le P(B)$.

5) $P(A)\le 1$.

6) $P(A+B)\le P(A)+P(B)$.

Chứng minh. □

Tính chất 2 được mang tên là “chuyển qua biến cố đối”: Nếu thấy khó khăn khi tính xác suất trực tiếp, có thể sẽ dễ hơn nếu ta tính xác suất của biến cố đối. Các tính chất 3, 3a – 3d gọi là quy tắc cộng xác suất.

Hai tính chất 7, 8 sau đây - được gọi là tính chất liên tục của xác suất - phải dùng đến tiên đề IIIa.

7) Nếu $A_1, A_2,...$ là dãy tăng các biến cố: $A_n \in \mathcal{F}$, $A_1\subset A_2\subset ...$, thì $A=\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ và

$$P(A)=P\left(\bigcup_{n=1}^{\infty} A_i\right)=\lim_{n\to\infty} P(A_n).$$

8) Nếu $A_1, A_2,...$ là dãy giảm các biến cố: $A_n\in\mathcal{F}$, $A_1\supset A_2\supset ...$, thì $A=\bigcap_{i=1}^{\infty} A_n \in \mathcal{F}$ và

$$P(A)=P\left(\bigcap_{i=1}^{\infty} A_n\right)=\lim_{n\to\infty} P(A_n).$$

1.1.7. Suy diễn xác suất

Trong ứng dụng của lý thuyết xác suất, ta hay gặp vấn đề sau:

Giả sử bằng cách nào đó, thông qua những quan sát của quá khứ, chúng ta biết được rằng xác suất của biến cố A trong một thí nghiệm là $p=P(A)\in[0; 1]$. Ta có thể nói gì về sự xảy ra của biến cố A trong 1 lần thử đơn lẻ tiếp theo?

Về vấn đề này, chúng ta tách làm 3 trường hợp sau đây.

i) Trường hợp p khá gần 0. Một biến cố có xác suất rất nhỏ, thậm chí bằng không vẫn có thể xảy ra khi thực hiện phép thử. Tuy nhiên, người ta chấp nhận nguyên lý sau đây, gọi là **nguyên lý xác suất nhỏ:**

Một biến cố có xác suất rất nhỏ thì thực tế có thể coi rằng, biến cố đó sẽ không xảy ra trong một (hoặc một vài) phép thử tương lai.

Một biến cố có thể coi là có xác suất nhỏ tuỳ thuộc vào bài toán cụ thể. Ví dụ, xác suất để 1 chuyến bay chở khách bị nạn bằng 0,01 không thể coi là nhỏ. Trái lại, xác suất để tàu hoả đường dài về ga cuối chậm quá 15 phút bằng 0,05 lại coi là nhỏ và có thể xem tàu hoả như thế là đúng giờ.

Xác suất nhỏ thường được chọn trong khoảng $0,00001 \div 0,1$, ví dụ 0,001; 0,005; 0,01; 0,02; 0,05; 0,1.

ii) Trường hợp p khá gần 1. Tương tự người ta có **nguyên lý xác suất lớn** sau đây:

Nếu biến cố ngẫu nhiên có xác suất rất lớn, thì thực tế có thể coi rằng, biến cố đó sẽ xảy ra trong một (hoặc một vài) phép thử tương lai.

iii) Trường hợp p khá xa 0 và 1.

Ví dụ $p(A) = 0,6$. (xem [1])

§1.2. XÁC SUẤT ĐIỀU KIỆN (1 tiết)

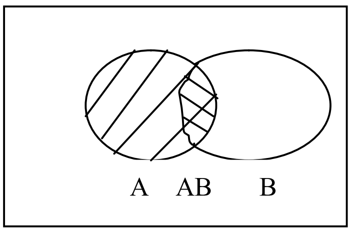
1.2.1. Xác suất điều kiện

Định nghĩa. Cho trước hai biến cố A, B với $P(A) > 0$. Xác suất của biến cố B tính trong điều kiện biến cố A đã xảy ra được gọi là xác suất điều kiện của biến cố B với điều kiện A, ký hiệu là P(B|A), xác định bởi

$$P(B|A) = \frac{P(AB)}{P(A)}, \qquad (P(A) > 0). \tag{1.2.1}$$

Mô tả xác suất điều kiện bằng lược đồ Venn cho ở Hình 1.3. Xét thí nghiệm gieo ngẫu nhiên 1 điểm trên miền G và giả sử đã biết điểm đó rơi vào miền A. Khi đó, khả năng điểm đó rơi vào miền B là

$$\frac{\text{diện tích miền } A \cap B}{\text{diện tích miền } A}$$



Hình 1.3

Mô tả bằng tần suất. Ký hiệu n_A, n_B, n_{AB} lần lượt là số lần xảy ra biến cố A trong loạt n phép thử với n đủ lớn.

Xem rằng $P(A) = \frac{n_A}{n}$; $P(B) = \frac{n_B}{n}$; $P(AB) = \frac{n_{AB}}{n}$.

$$\Rightarrow P(B|A) = \frac{P(AB)}{P(A)} = \frac{n_{AB}/n}{n_A/n} = \frac{n_{AB}}{n_A}.$$

Có thể kiểm tra định nghĩa này thoả mãn các tiên đề I, II, III (hoặc III_a) của xác suất, do đó, nó cũng là một xác suất. Vì thế nó có các tính chất của xác suất thông thường; ví dụ, $P(\overline{B}|A) = 1 - P(B|A)$. Sau đây là một số tính chất trực tiếp suy từ định nghĩa.

Định lý 1.1 *(Định lý nhân xác suất).*

Nếu $P(B) \neq 0$ thì $P(AB) = P(A \mid B)P(B)$. (1.2.2)

Định lý 1.2 *(Định lý nhân xác suất tổng quát).*

$$P(A_1A_2...A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2)$$
$$... P(A_n|A_1A_2...A_{n-1}). \quad (1.2.3)$$

Chứng minh (1.2.3) theo quy nạp.

Ví dụ 1.14*.* Gieo đồng thời 2 con xúc xắc. Tính xác suất để tổng số nốt trên 2 con xúc xắc ≥ 10 biết rằng ít nhất 1 con ra nốt 5.

Giải. A = {ít nhất 1 con ra nốt 5}; B = {tổng số nốt ≥ 10}.

$$P(A) = 1 - P(\overline{A}) = 1 - (5/6)^2 = 11/36.$$
$$AB = \{(5;6); (6;5); (5;5)\} \Rightarrow P(AB) = 3/36$$
$$\Rightarrow P(B|A) = P(AB)/P(A) = 3/11. \qquad \#$$

Ví dụ 1.15. Trong một hộp có 3 trục loại I và 7 trục loại II. Người thợ lắp máy rút ngẫu nhiên một chiếc sau đó rút ngẫu nhiên chiêc thứ hai. Tính xác suất để chiếc thứ nhất là trục loại 1 còn chiếc thứ hai là trục loại II.

Giải. Đặt A = {Chiếc thứ nhất loại I}; B = {Chiếc thứ hai loại II}.

(Ta phải hiểu A là biến cố chiếc thứ nhất rút được là trục loại I, còn chiếc thứ hai bất kỳ, loại nào cũng được…). Xác suất cần tìm là

$$P(AB) = P(A)P(B|A) = (3/10)(7/9) = 7/30 \approx 0,233.$$

Độc giả cũng có thể giải bằng cách sử dụng định nghĩa xác suất cổ điển khi xét thí nghiệm rút 2 phần tử có thứ tự từ 10 phần tử . #

1.2.2. Công thức xác suất toàn phần, công thức Bayes

Chúng ta nhắc lại (từ Ví dụ 1.11) rằng, nhóm các biến cố được gọi là đầy đủ nếu chúng xung khắc từng đôi và hợp của chúng là biến cố chắc chắn.

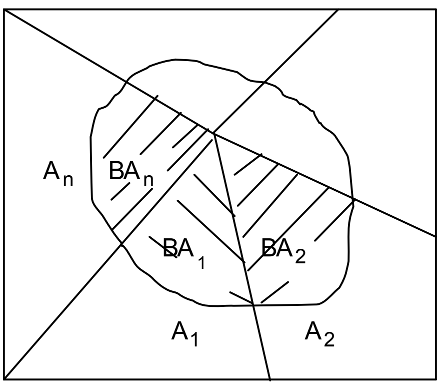
Định lý 1.3 *(Công thức xác suất toàn phần).* Giả sử $A_1, A_2, ..., A_n$ là nhóm đầy đủ các biến cố còn B là biến cố bất kỳ. Khi đó:

$$P(B) = P(B|A_1)P(A_1) + ... + P(B|A_n)P(A_n). \quad (1.2.4)$$

Chứng minh. Ta có

$$B = BS = B(A_1 + ... + A_n) = BA_1 + ... + BA_n.$$

Các biến cố $A_1, A_2, ..., A_n$ là xung khắc từng đôi nên $BA_1, ..., BA_n$ cũng xung khắc từng đôi (xem lược đồ Venn ở Hình 1.4).



Hình 1.4. Sự phân chia biến cố B thành các biến cố xung khắc.

Từ Định lý cộng và Định lý nhân ta được:

$$P(B) = P(BA_1) + \ldots + P(BA_n)$$

$$= P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_n)P(A_n) \qquad \square$$

***Định lý 1.4** (Công thức Bayes)*

Nếu $P(B) > 0$ và nhóm các biến cố $A_1, A_2, \ldots, A_n$ là đầy đủ thì

$$P(A_i \mid B) = \underbrace{\frac{P(B \mid A_i)P(A_i)}{P(B)}}_{\text{Biet } P(B)} = \underbrace{\frac{P(B \mid A_i)P(A_i)}{P(B \mid A_1)P(A_1) + \ldots + P(B \mid A_n)P(A_n)}}_{\text{Chua biet } P(B)} \qquad (1.2.5)$$

Chứng minh. Theo định nghĩa,

$$P(A_i \mid B) = \frac{P(BA_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{P(B)}.$$

Chỉ việc thay P(B) ở (1.2.4) vào mẫu ở vế phải . $\square$

Lưu ý. Nếu phép thử gồm hai giai đoạn thì các biến cố liên quan đến giai đoạn đầu thường được xem xét để lập nên nhóm đầy đủ các biến cố.

Ví dụ 1.16. Có 3 hộp bề ngoài giống hệt nhau. Các hộp chứa lần lượt 10, 15, 20 sản phẩm và mỗi hộp đều có 5 phế phẩm. Lấy ngẫu nhiên 1 hộp, từ đó rút ngẫu nhiên 1 sản phẩm.

a) Tính xác suất lấy được chính phẩm.

b) Kiểm tra thì thấy sản phẩm lấy được đúng là chính phẩm. Tính xác suất để sản phẩm đó được rút từ hộp thứ nhất.

Giải. a) Đặt $H_i = \{$Sản phẩm lấy được từ hộp thứ i$\}$, i = 1,2,3;

$A = \{$Rút được chính phẩm$\}$.

Rõ ràng H_1, H_2, H_3 là nhóm đầy đủ các biến cố; hơn nữa chúng đồng khả năng, vậy $P(H_1) = P(H_2) = P(H_3) = 1/3$. Theo công thức xác suất toàn phần

$$P(A) = P(A \mid H_1)P(H_1) + P(A \mid H_2)P(H_2) + P(A \mid H_3)P(H_3)$$

$$= \frac{1}{3}\frac{5}{10} + \frac{1}{3}\frac{10}{15} + \frac{1}{3}\frac{15}{20} = \frac{23}{36} = 0,639.$$

b) $P(H_1 \mid A) = \dfrac{P(A \mid H_1)P(H_1)}{P(A)} = \dfrac{6}{23} = 0,261.$ #

Ví dụ 1.17. Dây chuyền lắp ráp nhận các chi tiết từ 2 máy sản xuất ra. Trung bình máy thứ nhất cung cấp 60% chi tiết, máy thứ hai 40% . Khoảng 90% chi tiết do máy I và khoảng 40% chi tiết do máy II sản xuất ra đạt tiêu chuẩn. Lấy ngẫu nhiên 1 chi tiết từ dây chuyền thì thấy đạt yêu cầu. Tìm xác suất để chi tiết đó từ máy I sản xuất ra.

Giải. Hai biến cố H_i = {Chi tiết do máy I sản suất ra}, $(i = 1, 2)$ lập thành nhóm đầy đủ. Đặt A = {Chi tiết lấy ra đạt tiêu chuẩn} ...

Theo công thức Bayes,

$$P(H_1 \mid A) = \frac{P(A \mid H_1)P(H_1)}{P(A \mid H_1)P(H_1) + P(A \mid H_2)P(H_2)} = 0,771. \ \#$$

Ví dụ 1.18. Có 2 lồng chuột thí nghiệm, lồng thứ nhất có 10 con chuột đực và 15 con chuột cái, lồng thứ II có 8 con chuột đực và 7 con chuột cái. Bắt 1 con từ lồng I đưa sang lồng II; sau đó bắt 1 con từ lồng II thì được con chuột đực. Tính xác suất để con bắt được này từ lồng I.

Giải. Đặt A_i = {Bắt lần hai được chuột từ lồng i}, i = 1, 2; A_1, A_2 là nhóm đầy đủ. Lại đặt B = {Bắt lần hai được chuột đực}.

$$P(A_1|B) = \frac{P(B|A_1).P(A_1)}{P(B|A_1).P(A_1) + P(B|A_2).P(A_2)} = 0.769\ .\#$$

§1.3. SỰ ĐỘC LẬP (1 tiết)

1.3.1. Sự độc lập của 2 biến cố

Định nghĩa. Ta gọi hai biến cố A và B là độc lập nếu

$$P(AB) = P(A)P(B). \quad (1.3.1)$$

Rõ ràng là với biến cố A bất kỳ thì

$$\begin{aligned} &P(SA) = P(S)P(A); \\ &P(\varnothing A) = P(\varnothing)P(A). \end{aligned} \quad (1.3.2)$$

Vậy biến cố chắc chắn S và biến cố trống ∅ độc lập với biến cố bất kỳ.

Hai tính chất sau dễ dàng kiểm chứng.

Tính chất

a) Giả sử $P(B) \neq 0$, A và B độc lập $\Leftrightarrow P(A|B) = P(A)$.

Giả sử $P(\overline{B}) \neq 0$, A và B độc lập $\Leftrightarrow P(A|\overline{B}) = P(A)$.(1.3.3)

b) Giả sử $P(A) \neq 0$, A và B độc lập $\Leftrightarrow P(B|A) = P(B)$;

Giả sử $P(\overline{A}) \neq 0$, A và B độc lập $\Leftrightarrow P(B|\overline{A}) = P(B)$. (1.3.4)

Ý nghĩa. Tính chất (1.3.3), (1.3.4) nói lên rằng, nếu 2 biến cố là độc lập thì sự xuất hiện hay không của biến cố này không ảnh hưởng đến khả năng xuất hiện của biến cố kia. Thực tế, tiêu chuẩn trực giác này dùng để xét xem 2 biến cố đã cho có độc lập với nhau hay không.

Hệ quả. Nếu A và B là hai biến cố độc lập thì A và $\overline{B}$; $\overline{A}$ vμ B; $\overline{A}$ vμ $\overline{B}$ cũng là những cặp biến cố độc lập.

Ví dụ 1.19. Hai chị A và B cùng đến nhà hộ sinh để sinh con. Đặt

A = {Chị A sinh con trai}; B = {Chị B sinh con trai}.

Tìm xác suất cả hai chi đều sinh con trai. Rõ ràng dù A xảy ra hay $\overline{A}$ xảy ra thì khả năng sinh con trai của chị B vẫn không bị ảnh hưởng. Vậy ta coi hai biến cố A và B là độc lập, hơn nữa coi là đồng khả năng. Từ đó khả năng cả hai chị sinh con trai là

$$P(AB) = P(A)P(B) \approx (1/2)(1/2) = 1/4. \qquad \#$$

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	**Bài tập về nhà cho cả chương I** Tài liệu [1]: 1(2 – <u>3</u> – 5 – <u>7</u> – 9 – <u>10</u> - <u>11</u>–13 – <u>15</u> – <u>17</u> – <u>18</u> –<u>19</u> - <u>20</u> – <u>21</u> – <u>22</u> – 23 – <u>24</u> - <u>27</u> -29). Tài liệu [2]: Tr 35-38: 6, 9, 10, 12, 13, 15, 21, 25, 29, 30, <u>33</u>

	(sửa 10% thành 7%).
Tài liệu	Tài liệu [1], tr ....

Bài giảng 2: Biến cố và xác suất của biến cố (tiếp)

Chương, mục: 1

Tiết thứ: 5-8 Tuần thứ: 2

Mục đích, yêu cầu:

- Nắm được công thức Bernoulli và một số biến dạng của nó.
- Biết cách vận dụng lý thuyết để làm bài tập.

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§1.3. Sự độc lập (tiếp) : Phép thử lặp và công thức Bernoulli

Bài tập xác suất của biến cố, xác suất điều kiện

1.3.2. Sự độc lập của n biến cố

Định nghĩa.

Tính chất. Nếu $A_1, A_2, ..., A_n$ là những biến cố độc lập thì

$$P(A_1...A_n) = P(A_1)...P(A_n). \qquad (1.3.5)$$

Ví dụ 1.20. Một xí nghiệp có 3 ô tô hoạt động độc lập. Xác suất để trong ngày các ô tô này bị hỏng lần lượt là 0,1; 0,2; 0,3. Tìm xác suất để trong ngày có:

i) đúng 1 ô tô bị hỏng; ii) ít nhất 1 ô tô bị hỏng.

Giải. Đặt $A_i = \{\text{¤ t« thø i bÞháng trong ngµy}\}$;

$A = \{\text{Cã ®óng 1 « t« bÞháng}\}$;

$H = \{\text{Cã Ýt nhÊt 1 « t« bÞháng}\}$.

i) Ta có $A = A_1\overline{A_2}\,\overline{A_3} + \overline{A_1}\,A_2\,\overline{A_3} + \overline{A_1}\,\overline{A_2}A_3$.

Các biến cố ở vế phải là xung khắc từng đôi, mỗi số hạng là tích của các biến cố độc lập. Từ đó

$$P(A) = P(A_1\overline{A_2}\,\overline{A_3}) + P(\overline{A_1}\,A_2\,\overline{A_3}) + P(\overline{A_1}\,\overline{A_2}\,A_3)$$

$$= P(A_1)P(\overline{A_2})P(\overline{A_3}) + P(\overline{A_1})P(A_2)P(\overline{A_3}) + P(\overline{A_1})P(\overline{A_2})P(A_3) = 0,398$$

ii) Đặt $B = \{\text{Cã 2 « t« bÞháng}\}$; $C = \{\text{Cã 3 « t« bÞháng}\}$.

Các biến cố A, B, C xung khắc, vậy

$$D = A + B + C = A + (A_1A_2\overline{A_3} + A_1\overline{A_2}A_3 + \overline{A_1}A_2A_3) + A_1A_2A_3.$$

Giống như trên ta tính được P(D) = 0,496.

Nhận xét. Để tính P(D) ta có thể chuyển qua biến cố đối như sau.

$$\overline{D} = \{\text{Trong ngµy c¶ 3 « t« ®Òu tèt}\} = \overline{A_1}\,\overline{A_2}\,\overline{A_3}.$$

$$P(\overline{D}) = P(\overline{A_1})P(\overline{A_2})P(\overline{A_3}) = 0,9.0,8.0,7 = 0,504$$

$$\Rightarrow P(D) = 1 - P(\overline{D}) = 0,496. \qquad \#$$

1.3.3. Dãy các phép thử Bernoulli

Định nghĩa. Đối với thí nghiệm ngẫu nhiên nào đó chúng ta thực hiện n lần thử lặp lại. Chúng ta gọi dãy các phép thử này là dãy các phép thử Bernoulli nếu thoả mãn các điều kiện sau:

i) Đây là dãy các phép thử độc lập, nghĩa là kết quả của mỗi phép thử không phụ thuộc vào kết quả của các phép thử khác.

ii) Biến cố A xảy ra với xác suất p như nhau ở phép thử thứ i bất kỳ.

Nếu biến cố A xảy ra ở phép thử thứ i, ta nói phép thử thứ i thành công. Trái lại, nếu nó không xảy ra ở phép thử thứ i, ta nói phép thử này thất bại.

Định lý 1.5 *(Công thức Bernoulli).* Xác suất để biến cố A xuất hiện đúng k lần trong dãy n phép thử Bernoulli, ký hiệu là $P_n(k)$, hay đầy đủ hơn $P_n(k,p)$, được cho bởi công thức

$$P_n(k) = P_n(k;p) = C_n^k p^k (1-p)^{n-k},\ \ k = 0,1,...,n. \quad (1.3.6)$$

Chứng minh. Đặt $A_i = \{\text{BiÕn cè A x¶y ra ë lÇn thö thø i}\}$;

$B = \{\text{BiÕn cè A x¶y ra ®óng k lÇn}\}$. Thế thì

$$B = A_1A_2...A_{k-1}A_k\overline{A_{k+1}}...\overline{A_n} + ... + A_1A_2...A_{k-1}\overline{A_k}A_{k+1}\overline{A_{k+2}}...\overline{A_n}$$
$$+ ... + \overline{A_1}\,\overline{A_2}...\overline{A_{n-k}}A_{n-k+1}...A_n. \quad (1.3.7)$$

Mỗi số hạng ở vế phải (1.3.7) là tích của các biến cố độc lập bao gồm k thừa số A (với chỉ số) và $n-k$ thừa số $\overline{A}$ (với chỉ số); theo Định lý nhân và từ giả thiết, nó sẽ có xác suất $p^k(1-p)^{n-k}$. Mỗi số hạng ứng với 1 cách xếp k chữ cái A vào n chỗ, vậy có cả thảy C_n^k số hạng. Các số hạng là những biến cố xung khắc. Từ đó

$$P(B) = P(A_1A_2...A_{k-1}A_k\overline{A_{k+1}}...\overline{A_n}) + ... + P(\overline{A_1}\,\overline{A_2}...\overline{A_{n-k}}A_{n-k+1}...A_n)$$
$$= p^k(1-p)^{n-k} + ... + p^k(1-p)^{n-k} = C_n^k p^k(1-p)^{n-k}. \quad \square$$

Ngoài xác suất $P_n(k)$, người ta cũng hay xét

$$P_n(k_1 \div k_2) = \sum_{k=k_1}^{k_2} P_n(k) \quad (1.3.8)$$

là xác suất xảy ra biến cố A với số lần từ k_1 ®Õn k_2.

Ví dụ 1.22. Bắn 5 phát súng vào mục tiêu, xác suất trúng đích của mỗi phát là 0,2. Để phá huỷ mục tiêu cần từ 3 phát trúng đích trở lên. Tính xác suất phá huỷ mục tiêu.

Giải. Xem như chúng ta đã thực hiện dãy 5 phép thử độc lập. Biến cố mục tiêu bị phá huỷ là biến cố có 3 phát trúng đích trở lên. Từ đó

$$P = P_5(3\div 5; 0,2) = P_5(3;0,2) + P_5(4;0,2) + P_5(5;0,2)$$
$$= C_5^3 0,2^3 0,8^2 + C_5^4 0,2^4 0,8^1 + C_5^5 0,2^5 0,8^0 = 0,057.$$

Cũng có thể tra bảng: $P = 0,051 + 0,006 + 0,000 = 0,057$. #

Ví dụ 1.23. Gieo ngẫu nhiên n điểm trên khoảng (0; T). Xác suất để có đúng k điểm trên khoảng $(a;b) \subset (0;T)$ là bao nhiêu? Xét trường hợp $k = 0;\ n$.

Giải. Xem như ta thực hiện n phép thử độc lập, ở đó phép thử đơn là gieo 1 lần 1 điểm, biến cố A là điểm đơn rơi vào khoảng (a; b) với xác suất $p = (b-a)/T$. Như vậy biến cố cần tính xác suất là biến cố {A xảy ra đúng k lần}. Theo công thức Bernoulli,

$$P = C_n^k p^k (1-p)^{n-k}, \text{ ví i } p(b\text{-}a)/T.$$

Khi $k = 0$ th× $P = (1-p)^n$; $k = n$ th× $P = p^n$. #

BÀI TẬP: Xác suất biến cố (1 tiết)

Xác suất điều kiện (2 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 3: Biến ngẫu nhiên

Chương, mục: 2

Tiết thứ: 9-12 Tuần thứ: 3

Mục đích, yêu cầu:

- Thấy được nghiên cứu BNN là sự tiếp tục của biến cố.
- Tính được kỳ vọng, phương sai của các BNN liên tục, rời rạc

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

Bài tập chương 1 (1 tiết)

Chương 2. Biến ngẫu nhiên

§2.1. Biến ngẫu nhiên và luật phân bố

§ 2.2. Các đặc trưng số của biến ngẫu nhiên

BÀI TẬP: Sự độc lập (1 tiết)

Chương 2
BIẾN NGẪU NHIÊN

§2.1. BIẾN NGẪU NHIÊN VÀ LUẬT PHÂN BỐ (2tiết)

2.1.1. Mở đầu

Thông thường người ta hiểu biến ngẫu nhiên (BNN) (random variable) là đại lượng mà giá trị của nó phụ thuộc vào kết quả ngẫu nhiên của phép thử.

Định nghĩa. (Định nghĩa chính xác xem [1])

Nếu tập giá trị hữu hạn hay vô hạn đếm được thì BNN được gọi là rời rạc. Nếu tập giá trị lấp đầy một hoặc một số khoảng thì BNN được gọi là liên tục.

Thường người ta ký hiệu BNN bởi chữ cái in hoa: X, Y, Z,… hoặc có thêm chỉ số: $X_1, X_2, ...$

Như vậy, BNN không phải là biến số độc lập, nó là hàm số; hàm này xác định trên không gian các biến cố sơ cấp S.

Khi đó với bất kỳ $x, x_1, x_2 \in \mathbb{R}$ cho trước, các tập con sau đây của S:

$$(X = x) = \{\zeta : X(\zeta) = x\},$$

$$(X \le x) = \{\zeta : X(\zeta) \le x\},$$

$$(X > x) = \{\zeta : X(\zeta) > x\},$$

$$(x_1 < X \le x_2) = \{\zeta : x_1 < X(\zeta) \le x_2\},$$

$(X \in B) = \{\zeta : X(\zeta) \in B\}$, B - tập (đo được) tuỳ ý của $\mathbb{R}$

là những biến cố, muốn (và có thể) tính xác suất.

Đòi hỏi đo được của BNN mang tính chất toán học thuần tuý, có thể bỏ qua trong những bài toán thực tiễn.

Ví dụ 2.1. Tung con xúc xắc cân đối; mỗi nột trên mặt con xúc xắc được thưởng 10 USD. Đặt X bằng 10 lần số nốt trên mặt con xúc xắc, X là một BNN.

Ta mô tả kỹ hơn BNN này. Không gian các biến cố sơ cấp là $S=\{M_1,...,M_6\}$, M_i = {xuất hiện mặt có i nốt}. Đặt $X(M_1)=10$; $X(M_2)=20$; …; $X(M_6)=60$. Tập giá trị của X là {10; … ;60}; X là BNN rời rạc. #

Ví dụ 2.2. Số phế phẩm có trong lô hàng n sản phẩm. Không nói trước được số phế phẩm là bao nhiêu, đây là BNN rời rạc. Tập giá trị là {0; … ; n}. #

Ví dụ 2.3. Tuổi thọ của một linh kiện điện tử, đó là BNN liên tục; có thể, tập giá trị là [0; 10 000] (giờ). #

Việc nghiên cứu BNN là tổng quát hơn so với chỉ nghiên cứu biến cố thông thường.

2.1.2. Luật phân bố của biến ngẫu nhiên

Việc biết tập giá trị của BNN là quan trọng, song hai BNN có tập giá trị giống nhau lại có thể hoàn toàn khác nhau. Tập giá trị cho ta rất ít thông tin về BNN. Điều quan trọng là biết BNN nhận các giá trị có thể của nó với xác suất bao nhiêu.

Định nghĩa. Mối quan hệ giữa các giá trị có thể của BNN với xác suất tương ứng được gọi là luật phân bố của BNN ấy.

a) Luật phân bố của BNN rời rạc.

Định nghĩa. Giả sử $\{x_1, x_2, ...\}$ là tập giá trị của BNN X. Bộ số

$$p_1, p_2, ... \text{ víi } p_i = P(X = x_i),\ i = 1,2... \qquad (2.1.3)$$

được gọi là luật phân bố của BNN rời rạc X.

Rõ ràng rằng bộ số $p_1, p_2, ...$ thoả mãn điều kiện

$$p_i \geq 0;\ \sum_i p_i = 1 \text{ (tổng hữu hạn hay vô hạn).} \qquad (2.1.4)$$

Để thuận lợi, người ta sắp xếp bộ số $p_1, p_2, ...$ thành bảng: Dòng trên ghi các giá trị của BNN nhận (thường theo thứ tự tăng dần), dòng dưới ghi các xác suất tương ứng. Bảng này gọi là bảng phân bố xác suất (để đơn giản: bảng xác suất) của BNN rời rạc.

X	x_1	x_2	. . .	x_n . . .
P	p_1	p_2	. . .	p_n . . .

Ví dụ 2.6. Lập bảng xác suất của số nốt xuất hiện khi tung con xúc xắc.

Giải. Vì $p_i = P(X = i) = 1/6$ nên ta nhận được bảng

X	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

Nói chung, khi $p_1 = ... = p_n = 1/n$, X gọi là có phân bố đều rời rạc trên tập $\{x_1, ..., x_n\}$. Luật này được sử dụng khi thiếu thông tin về X; khi ấy, coi các giá trị mà nó có thể nhận là đồng khả năng. #

Ví dụ 2.7. Xác suất để một xạ thủ bắn trúng bia là 0,8. Xạ thủ bắn từng phát cho đến khi trúng bia. Lập bảng phân bố xác suất của số đạn cần bắn. Tính các xác suất

$$P\{-5 < X < 2,1\}; \quad P\{1,5 \le X \le 3\}.$$

Giải. Coi kết quả ở các lần bắn là độc lập. Số đạn X cần thiết có tập giá trị là $\{1,2,\dots\}$. Ta có

$$P(X = 1) = 0,8;\ P(X = 2) = 0,2 \cdot 0,8;\ \dots;\ P(X = i) = 0,2^{i-1}.0,8;\dots$$

Từ đó nhận được bảng xác suất

X	1	2	...	i	...
P	0,8	0,16	...	$0,2^{i-1}0,8$	...

Dễ kiểm tra điều kiện (2.1.4) thoả mãn. Từ bảng này ta có thể tính được mọi xác suất quan tâm. Chẳng hạn

$$P\{-5 < X < 2,1\} = P\{X = 1\} + P\{X = 2\} = 0,96;$$

$$P(1,5 \le X \le 3) = P\{X = 2\} + P\{X = 3\} = 0,192; \qquad \#$$

Cùng với bảng xác suất người ta cũng dùng hàm khối lượng xác suất (probability mass function), gọi tắt là hàm xác suất:

$$p_X(x) = P(X = x). \qquad (2.1.5)$$

b) Hàm phân bố.

Hàm phân bố (tên khác: hàm phân bố xác suất, hàm phân bố tích luỹ (cumulative distibution function)) của BNN X, ký hiệu $F_X(x)$, xác định bởi

$$F_X(x) = P(X \le x),\ x \in \mathbb{R}. \qquad (*) \qquad (2.1.6)$$

Nhận thấy rằng hàm phân bố được định nghĩa cho cả BNN rời rạc lẫn BNN liên tục. Sau đây là một số tính chất quan trọng của hàm phân bố.

Định lý 2.1. Hàm phân bố $F_X(x)$ của BNN X có các tính chất sau

i) $F_X(x)$ là hàm không giảm, tức là

$$F_X(x) \le F_X(y) \text{ khi } x < y. \qquad (2.1.7)$$

ii) $F_X(-\infty) \triangleq \lim\limits_{x \to -\infty} F_X(x) = 0; \qquad F_X(+\infty) \triangleq \lim\limits_{x \to +\infty} F_X(x) = 1.$ (2.1.8)

iii) $F_X(x)$ là hàm liên tục phải, tức là

$$F_X(x_0+) \triangleq \lim_{x \to x_0^+} F_X(x) = F_X(x_0),\quad x_0 \in \mathbb{R}. \qquad (2.1.9)$$

Chứng minh.

i) Từ chỗ $(X \le x) \subset (X \le y)$, do tính đơn điệu của xác suất nên

$$F_X(x) = P(X \le x) \le P(X \le y) = F_X(y).$$

ii) Xét hai dãy biến cố $(X \le -n); (X \le n)$. Thế thì $(X \le -n) \downarrow \varnothing;\ (X \le n) \uparrow S$ $(\text{khi } n \to \infty)$. Theo tính chất liên tục của xác suất (xem mục 1.1.6) ta được

$$F_X(-\infty) = \lim_{n \to \infty} F_X(-n) = \lim_{n \to \infty} P(X \le -n) = P(\varnothing) = 0;$$

$$F_X(+\infty) = \lim_{n \to \infty} F_X(n) = \lim_{n \to \infty} P(X \le n) = P(S) = 1.$$

iii) Giả sử $\{\varepsilon_n\}$ là dãy giảm đến 0 bất kỳ, thế thì $(X \le x_0 + \varepsilon_n) \downarrow (X \le x_0)$ (khi $n \to \infty$). Lại theo tính chất liên tục của xác suất suy ra

$$F_X(x_0 + \varepsilon_n) = P(X \le x_0 + \varepsilon_n) \downarrow P(X \le x_0) = F_X(x_0) \quad (n \to \infty).\square$$

Ngoài ra, hàm phân bố còn có các tính chất sau đây:

iv) $0 \le F_X(x) \le 1$

v) $P(a < X \le b) = F_X(b) - F_X(a)$ (2.1.10)

v') $P(a \le X \le b) = F_X(b) - F_X(a-)$

vi) $P(X > a) = 1 - F(a)$

vii) Nếu $F(a) = 0 \Rightarrow F(x) = 0, \ \forall x \le a.$

Công thức (2.1.10) có ý nghĩa rất quan trọng, nó cho phép tính xác suất của mọi biến cố quan tâm thông qua hàm phân bố.

c) Hàm phân bố của BNN rời rạc.

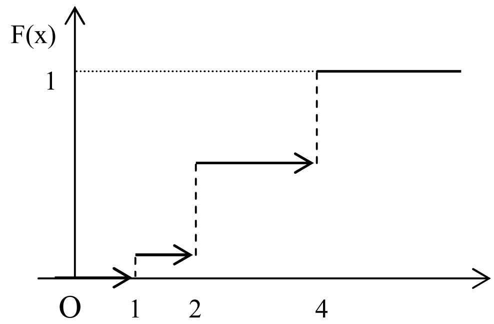
Ví dụ 2.8. Cho BNN rời rạc X có bảng xác suất như sau

X	1	2	4
P	0,1	0,5	0,4

Ta đi tìm hàm phân bố của BNN này.

$$x < 1 \Rightarrow F(x) = P(X < x) = P(\varnothing) = 0$$
$$1 \le x < 2 \Rightarrow F(x) = P(X = 1) = 0,1$$
$$2 \le x < 4 \Rightarrow F(x) = P(X = 1) + P(X = 2) = 0,1 + 0,5 = 0,6$$
$$4 \le x \Rightarrow F(x) = P(X = 1) + P(X = 2) + P(X = 4) = 1.$$

$$F(x) = \begin{cases} 0 & \text{với} \quad x < 1 \\ 0,1 & 1 \le x < 2 \\ 0,6 & 2 \le x < 4 \\ 1 & 4 \le x \end{cases}$$



Hình 2.2. Hàm phân bố của BNN trong Ví dụ 2.8

Nếu X là BNN rời rạc bất kỳ thì từ định nghĩa dễ suy ra

$$F(x) = \sum_{i:\, x_i \le x} p_i = \begin{cases} 0 & \text{với} \quad x < x_1 \\ p_1 & x_1 \le x < x_2 \\ p_1 + p_2 & x_2 \le x < x_3 \\ \cdot\ \cdot\ \cdot\ \cdot \end{cases} \qquad (2.1.11)$$

Đây là hàm bậc thang, gián đoạn tại x_i thuộc tập giá trị với bước nhảy p_i tương ứng: $F(x_i) - F(x_i -) = p_i$. #

d) Hàm phân bố của BNN liên tục, hàm mật độ.

Bây giờ ta đưa ra định nghĩa chính xác hơn về BNN liên tục.

Định nghĩa. Biến ngẫu nhiên X được gọi là liên tục nếu hàm phân bố F(x) của nó liên tục trên $\mathbb{R}$, hàm F(x) khả vi có thể trừ ra tại một số hữu hạn hoặc đếm được điểm. Đạo hàm của hàm phân bố gọi là hàm mật độ:

$$f(x) = \frac{dF(x)}{dx} = F'(x). \qquad (2.1.12)$$

Định lý 2.2

Giả sử X là BNN liên tục với hàm mật độ $f_X(x)$. Khi đó

i) $f_X(x)$ là hàm không âm, tức là

$$f_X(x) \ge 0, \ \forall x \in \mathbb{R}. \qquad (2.1.13)$$

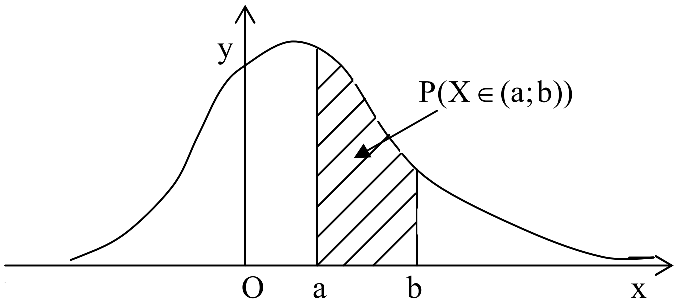
ii) $f_X(x)$ khả tích trên $\mathbb{R}$ và tích phân của nó bằng 1:

$$\int_{-\infty}^{\infty} f_X(x)dx = 1. \qquad (2.1.14)$$

iii) $F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \ x \in \mathbb{R}. \qquad (2.1.15)$

iv) Xác suất để X nhận một giá trị cụ thể bất kỳ bằng không :

$$P(X = a) = 0, \quad a \in \mathbb{R}; \qquad (2.1.16)$$



Hình 2.3. Hàm mật độ và tính xác suất từ hàm mật độ

Hơn nữa

$$P(a < X < b) = P(a \le X \le b) = \int_a^b f_X(x)dx, \ a, b \in \mathbb{R}. \ (2.1.17)$$

Nhận xét. Xác suất để BNN X nhận giá trị trong khoảng (a; b) nào đó bằng diện tích hình thang cong giới hạn bởi đường cong mật độ và các đường thẳng $x = a, \ x = b, \ y = 0$ (Hình 2.3).

Mở rộng: Nếu A là tập (đo được) bất kỳ của $\mathbb{R}$ thì

$$P(X \in A) = \int_A f_X(x)dx \qquad (2.1.18)$$

Ví dụ 2.9. BNN X gọi là có phân bố đều trên (a; b) nếu

$$f_X(x) = \begin{cases} c & \text{víi} \quad a < x < b \\ 0 & \quad \text{tr¸i l¹i.} \end{cases}$$

Chúng ta tìm hằng số c và hàm phân bố F(x). Từ (2.1.14) suy ra

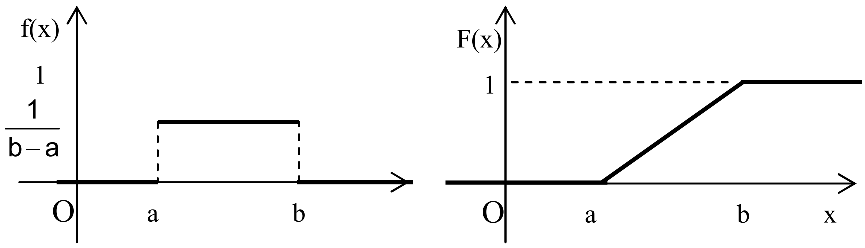
$$\int_{-\infty}^{\infty} f(x)dx = \int_a^b c dx = c(b-a) = 1 \Rightarrow c = \frac{1}{b-a}.$$

Với $x < a$ thì $F(x) = 0$; với $x > a$ thì $F(x) = 1$.

Với $x \in [a;b]$ thì $F(x) = \int\limits_{-\infty}^{x} f(x)dx = \int\limits_{a}^{x} \frac{1}{b-a}dx = \frac{x-a}{b-a}$

là hàm bậc nhất. Đồ thị của hàm mật độ và phân bố cho ở Hình 2.4.

Trường hợp đặc biệt khi $(a;b) = (0;1)$, lúc đó $F(x) = x$ và $f(x) = 1$ trên (0; 1).



Hình 2.4. Hàm mật độ và hàm phân bố của phân bố đều trên (a; b)

Bây giờ cho [c; d] là đoạn bất kỳ trong [a; b], theo (2.1.18) thì:

$$P\{c \le X \le d\} = \frac{d-c}{b-a} = \frac{\text{Độ dài đoạn } [c;d]}{\text{Độ dài đoạn } [a;b]}$$

Điều này phù hợp với định nghĩa xác suất hình học trên $\mathbb{R}$. #

Ví dụ 2.10. BNN X có hàm mật độ

$$f(x) = \begin{cases} a\cos x & \text{nếu} \quad -\pi/2 < x < \pi/2 \\ 0 & \text{trái lại.} \end{cases}$$

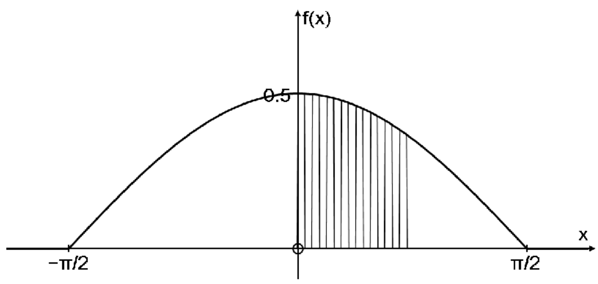
a) Tìm a và vẽ đồ thị mật độ thu được.

b) Tính F(x), từ đó tính $P(0 < X < \pi/4)$.

Giải. a) $1 = \int\limits_{-\infty}^{\infty} f(x)dx = \int\limits_{-\pi/2}^{\pi/2} a\cos x dx = 2a \Rightarrow a = \frac{1}{2}$.

b) $F(x) = \int\limits_{-\infty}^{x} f(t)dt = \begin{cases} 0 & \text{với} \quad x \le -\pi/2 \\ (1/2)(1+\sin x) & -\pi/2 < x \le \pi/2 \\ 1 & \pi/2 < x. \end{cases}$

$P(0 < X < \pi/4) = F(\pi/4) - F(0) = \sqrt{2}/4.$ #



Hình 2.5. Mật độ của BNN ở Ví dụ 2.10

§ 2.2. CÁC ĐẶC TRƯNG SỐ CỦA BIẾN NGẪU NHIÊN (1 tiết)

2.2.1. Các đặc trưng về giá trị trung tâm của BNN

a) Kỳ vọng

Định nghĩa. Kỳ vọng (hay giá trị trung bình) của BNN X, ký hiệu là E[X], được xác định như sau:

* Nếu X là BNN rời rạc với $p_i = P(X = x_i),\ i = 1, 2, \ldots$ thì

$$E[X] = \sum_{i=1}^{\infty} x_i p_i. \qquad (2.2.1)$$

(Chuỗi trở thành tổng hữu hạn nếu tập giá trị của X là hữu hạn).

* Nếu X là BNN liên tục với hàm mật độ $f_X(x)$ thì

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx. \qquad (2.2.2)$$

Kỳ vọng thể hiện giá trị trung tâm của BNN, nó là đặc trưng quan trọng nhất, dễ làm việc nhất.

Định nghĩa. BNN X được gọi là quy tâm nếu kỳ vọng của nó bằng không: $E[X] = 0$. Đối với BNN X bất kỳ, BNN $Y = X - E[X]$ là quy tâm, gọi là BNN quy tâm hoá của X.

Ví dụ 2.11. Tính kỳ vọng của BNN X và Y khi:

i. X có bảng xác suất

X	1	3	4
P	0,1	0,5	0,4

ii) Y có hàm mật độ $f(x) = \begin{cases} (3/4)(x^2 + 2x), & \text{ví i} \quad x \in (0; 1) \\ 0, & \text{tr¸ i l¹i.} \end{cases}$

Giải. $\quad$ i) $E[X] = 1.0,1 + 3.0,5 + 4.0,4 = 3,2.$

ii) $E[Y] = \int_{-\infty}^{\infty} x f(x)dx = \int_0^1 x \frac{3}{4}(x^2 + 2x)dx = \ldots = \frac{11}{16}.$ #

Định lý 2.3. Kỳ vọng của BNN có các tính chất sau đây

i) $\quad E[C] = C$

$$E[aX] = aE[X],\ a \in \mathbb{R}. \qquad (2.2.4)$$

ii) Nếu X và Y là những BNN có kỳ vọng thì tổng $X + Y$ cũng có kỳ vọng và

$$E[X + Y] = E[X] + E[Y]. \qquad (2.2.5)$$

iii) Nếu X và Y là hai BNN độc lập và có kỳ vọng thì tích XY cũng có kỳ vọng và

$$E[XY] = E[X].E[Y]. \qquad (2.2.6)$$

iv) Giả sử $\varphi(x)$ là hàm số thông thường nào đó sao cho $Y = \varphi(X)$ là BNN có kỳ vọng. Khi đó:

X rời rạc: $$E[Y] = \sum_{i=-\infty}^{\infty} \varphi(x_i)p_i; \qquad (2.2.7)$$

X liên tục: $$E[Y] = \int_{-\infty}^{\infty} \varphi(x) f_X(x)dx. \qquad (2.2.8)$$

Nhận xét. Các tính chất i), ii) gọi là tính chất tuyến tính của kỳ vọng, giống với tích chất tuyến tính của tổng hay tích phân. Tính chất iii) liên quan đến tính độc lập của các BNN sẽ nói đến kỹ hơn ở mục 3.2.3. Tất nhiên, để tính kỳ vọng của $Y=\varphi(X)$, ta có thể tìm phân bố của Y (bảng xác suất hay hàm mật độ) rồi áp dụng định nghĩa. Tuy nhiên áp dụng (2.2.7) hay (2.2.8) là thuận lợi hơn nhiều.

Ví dụ 2.12. Tính $E[X^2]$ với

X	–3	1	3	5	6
P	0,1	0,2	0,3	0,2	0,2

Giải. Cách I. BNN $Y=X^2$ có bảng phân bố

Y	1	9	25	36
P	0,2	0,4	0,2	0,2

Từ đó, $E[Y]=1.0,2+9.0,4+25.0,2+36.0,2=16$.

Cách II. Sử dụng (2.2.7) thì

$$E[X^2]=(-3)^2.0,1+1^2.0,2+3^2.0,3+...+6^2.0,2=16.\#$$

Ví dụ 2.13. Giả sử X là BNN với hàm mật độ f(x). Khi đó

$$E\left[\frac{X^3-1}{X^2+2}\right]=\int_{-\infty}^{\infty}\frac{x^3-1}{x^2+2}f(x)dx;$$

$$E\left[\ln(X^2)\right]=\int_{-\infty}^{\infty}\ln(x^2)\,f(x)dx. \qquad \#$$

b. Mốt. Mốt của BNN X, ký hiệu là Mod[X], là giá trị mà tại đó xác suất tương ứng hay hàm mật độ đạt giá trị cực đại, cụ thể là:

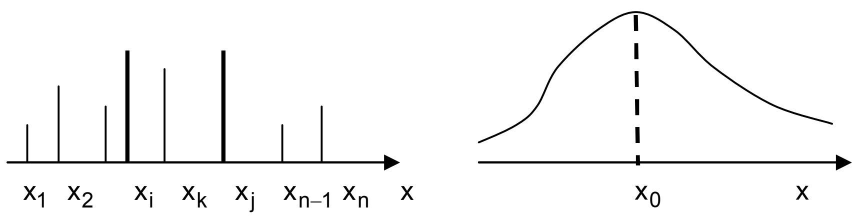
X rời rạc, $p_i=P(X=x_i),\ i=1,2,...;$

$$x_k=Mod[X]\Leftrightarrow p_k=Max(p_1,p_2,...).$$

X liên tục với hàm mật độ f(x);

$$x_0=Mod[X]\ \Leftrightarrow\ f(x_0)=Max f(x).$$

Như vậy một BNN có thể có một hoặc một số mốt. Nếu chỉ có đúng 1 mốt thì BNN được gọi là đơn mốt.



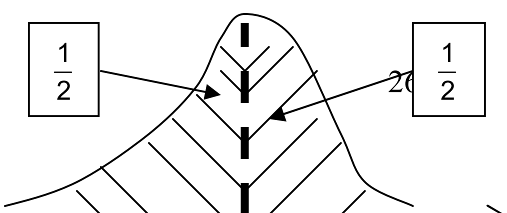
Hình 2.6. Mốt của BNN rời rạc (a) và của BNN liên tục (b).

c) Trung vị (median).

Trung vị (median) của BNN X, ký hiệu là Med[X], xác định bởi

X rời rạc: $Med[X]=x_i$ sao cho $\begin{cases}P(X\le x_i)\ge 1/2;\\ P(X\ge x_i)\le 1/2.\end{cases}$

X liên tục: $Med[X]=x_0\ \Leftrightarrow\ \int_{-\infty}^{x_0}f_X(x)dx\ge\frac{1}{2};\ \int_{x_0}^{\infty}f_X(x)dx\ge\frac{1}{2}.$



Hình 2.7. Trung vị của BNN liên tục.

Nói chung, $\mathrm{Med}[X] \neq E[X]$. Tuy nhiên, khi hàm xác suất hay hàm mật độ đối xứng qua một trục nào đó thì trung vị trùng với kỳ vọng. Trung vị đại diện rất tốt cho giá trị trung tâm của BNN, song tính toán nó không được thuận lợi như với kỳ vọng.

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	**Bài tập về nhà cho cả chương II** Tài liệu [1]: 2(1 - 2 –3 - 4– 5- 6 - 7 - 8 – 9 - 10 - 11 –12- 14 – 16-17 - 18-21- 23- 26 - 27- 30-32). Tài liệu [2]: Tr 76-78: 2, 4, 8 (sửa x thành $\|x\|$), 10.
Tài liệu	Tài liệu [1], tr ....

Bài giảng 4: Biến ngẫu nhiên (tiếp)

Chương, mục: 2

Tiết thứ: 13-16 Tuần thứ: 4

Mục đích, yêu cầu:

- Tính được phương sai của BNN
- Nắm chắc về phân bố nhị thức, đều.
- Làm được bài tập

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§ 2.2. Các đặc trưng số của biến ngẫu nhiên (tiếp)

§2.3. Một số luật phân bố quan trọng

Bài tập về biến ngẫu nhiên

§ 2.2. CÁC ĐẶC TRƯNG SỐ CỦA BIẾN NGẪU NHIÊN (tiếp - 1 tiết)

2.2.2. Các đặc trưng về độ phân tán của biến ngẫu nhiên

a) Phương sai.

Định nghĩa. Phương sai của BNN X, ký hiệu là V[X] (có tài liệu ghi là Var[X] hay DX), xác định bởi

$$V[X] = E[(X - E[X])^2]. \qquad (2.2.9)$$

Khai triển vế phải của (2.2.9) ta suy ngay ra

$$V[X] = E[X^2] - (E[X])^2. \qquad (2.2.10)$$

Sử dụng (2.2.10) cũng như tính chất iv) của kỳ vọng nhận được các công thức sau đây rất thuận lợi khi tính phương sai.

X rời rạc: $$V[X] = \sum_{i=1}^{\infty}(x_i - E[X])^2 p_i$$

$$= \sum_{i=1}^{\infty} x_i^2 p_i - (E[X])^2. \qquad (2.2.11)$$

X liên tục: $$V[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f_X(x)dx$$

$$= \int_{-\infty}^{\infty} x^2 f_X(x)dx - (E[X])^2. \qquad (2.2.12)$$

Căn bậc hai của phương sai được gọi là độ lệch chuẩn, ký hiệu là σ_X:

$$\sigma_X = \sqrt{V[X]}. \qquad (2.2.13)$$

Định lý 2.4. Phương sai của BNN có các tính chất sau đây:

$V[C]=0;$

$V[aX]=a^2V[X];$

$V[X+C]=V[X]$ (a, C- hằng số) (2.2.14)

Nếu X và Y là hai BNN độc lập, có phương sai hữu hạn thì $X+Y$ cũng là BNN có phương sai hữu hạn và

$$V[X+Y]=V[X]+V[Y]. \qquad (2.2.15)$$

Hệ quả. Nếu hai BNN X,Y độc lập, có phương sai hữu hạn thì

$$V[X-Y]=V[X]+V[Y]. \qquad (2.2.16)$$

Nếu $X_1,...,X_n$ là các BNN độc lập, có phương sai hữu hạn thì

$$V[X_1+...+X_n]=V[X_1]+...+V[X_n]. \qquad (2.2.17)$$

b) Mômen.

$\nu_k[X]=E[X^k];$

$\mu_k[X]=E[(X-E[X])^k].$

c) Một số đặc trưng khác.

Độ lệch: $$Skew[X]=\frac{E(X-\mu)^3}{\sigma^3}=\frac{\mu_3}{\sigma^3}; \qquad (2.2.18)$$

Độ nhọn: $$Kurt[X]=\frac{E(X-\mu)^4}{\sigma^4}-3=\frac{\mu_4}{\sigma^4}-3. \qquad (2.2.19)$$

Độ rộng dải biến thiên. Nếu tập giá trị của X là $[m;M]$ thì:

m : giá trị nhỏ nhất, ký hiệu là Min[X];
M : giá trị lớn nhất, ký hiệu là Max[X];
[m; M] : dải biến thiên (hay tập giá trị);
$M-m$: độ rộng dải biến thiên.

§2.3. MỘT SỐ LUẬT PHÂN BỐ QUAN TRỌNG (1 tiết)

2.3.1. Luật phân bố nhị thức B(n;p)

BNN X được gọi là có phân bố nhị thức với tham số (n, p), ký hiệu $X \sim B(n;p)$, nếu

$$P(X=k)=C_n^k p^k (1-p)^{n-k},\ k=0,1,...,n. \qquad (2.3.1)$$

Khi $n=1$ chúng ta nhận được luật phân bố Bernoulli với

$P(X=1)=p;\ \ P(X=0)=1-p=q.$

Từ công thức Bernoulli (1.3.6) thì:

Số lần thành công (tức là số lần xuất hiện biến cố A) trong dãy n phép thử Bernoulli có phân bố nhị thức $B(n;p)$.

Định lý 2.5. Nếu $X \sim B(n;p)$ thì

$$\begin{cases} E[X]=np; \\ V[X]=npq. \end{cases} \qquad (2.3.2)$$

Chứng minh. Xét dãy phép thử Bernoulli. Gọi Y_i là số lần xuất hiện biến cố A trong phép thử thứ i. Thế thì $Y_1,...,Y_n$ là các BNN độc lập, cùng phân bố B(1; p). Ta có

$$E[Y_i] = 1.p + 0.(1-p) = p;$$

$$V[Y_i] = E[Y_i^2] - E^2[Y_i] = (1^2.p + 0^2.(1-p)) - p^2 = pq.$$

Đặt $Y = Y_1 + ... + Y_n$ ta được

$$E[Y] = E[Y_1] + ... + E[Y_n] = np;$$

$$V[Y] = V[Y_n] + ... + V[Y_n] = npq.$$

Ta biết rằng $Y \sim B(n; p)$, cùng luật phân bố với X. Vậy X có cùng kỳ vọng, phương sai như của Y. Chúng ta nhận được (2.3.2).□

2.3.2. Luật phân bố Poisson $P(\lambda)$

BNN X được gọi là có phân bố Poisson với tham số $\lambda > 0$, ký hiệu $X \sim P(\lambda)$, nếu

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad (k = 0, 1, 2, ...) \qquad (2.3.3)$$

Định lý 2.6. Nếu $X \sim P(\lambda)$ thì

1) $\begin{cases} E[X] = \lambda; \\ V[X] = \lambda. \end{cases} \qquad (2.3.4)$

2) Nếu $X \sim P(\lambda_1)$, $Y \sim P(\lambda_2)$; X vµ Y ®éc lËp th×

$$X + Y \sim P(\lambda_1 + \lambda_2). \qquad (2.3.5)$$

Một số ví dụ BNN có phân bố Poisson là:

Số lần gọi đến 1 tổng đài trong khoảng thời gian nào đó (trong 1 giờ, 1 ngày), số lần khách hàng đến nhà băng trong 1 giờ…

Nói chung, đầu vào một hệ phục vụ thường có phân bố Poisson.

2.3.3. Luật phân bố đều U(a; b)

a) Kỳ vọng, phương sai. BNN liên tục X gọi là có phân bố đều trên $(a; b)$, ký hiệu là $X \sim U(a; b)$, nếu hàm mật độ của nó có dạng

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{víi } x \in [a;\ b]; \\ 0 & \text{tr¸i l¹i.} \end{cases} \qquad (2.3.6)$$

Ở Ví dụ 2.9 ta đã xét hàm mật độ và hàm phân bố của X. Bây giờ ta tính kỳ vọng và phương sai của nó. Ta có

$$\begin{cases} E[X] = \dfrac{a+b}{2}; \\ V[X] = \dfrac{(b-a)^2}{12}. \end{cases} \qquad (2.3.7)$$

Thực vậy, $E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$

$$V[X] = E[X^2] - (E[X])^2 = \int_a^b x^2 \frac{1}{b-a} dx - \left((a+b)/2\right)^2 = \frac{(b-a)^2}{12}.$$

Đôi khi trong các kết luận thống kê ta hay sử dụng quy tắc sau:

Nếu biết rằng BNN X nhận giá trị trong khoảng (a; b) nào đó mà không biết thêm thông tin gì khác về X thì có thể xem mỗi giá trị có thể của X trong khoảng (a; b) là đồng khả năng; nói cách khác, X có phân bố đều trên (a; b).

b) Dãy số ngẫu nhiên.

Như vậy, bằng máy tính bất kỳ ta sinh ra dãy số ngẫu nhiên $\{x_1, x_2, ...\}$.

BÀI TẬP: Biến ngẫu nhiên và hàm phân bố (2tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 5: Biến ngẫu nhiên (tiếp)

Chương, mục: 2

Tiết thứ: 17-20 Tuần thứ: 5

Mục đích, yêu cầu:

- Nắm chắc về phân bố chuẩn
- Các phân bố T, F, Khi bình phương chỉ cần hiểu khái niệm phân vị.
- Làm được bài tập

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§2.3. Một số luật phân bố quan trọng (tiếp)

Bài tập về biến ngẫu nhiên (tiếp)

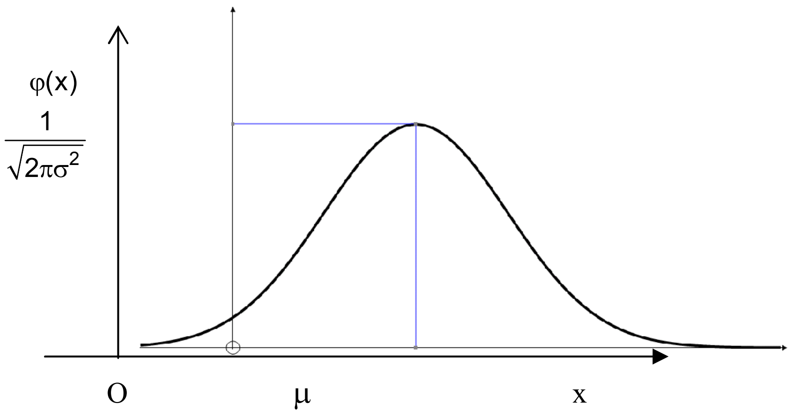
§2.3. MỘT SỐ LUẬT PHÂN BỐ QUAN TRỌNG (tiếp - 1 tiết)

2.3.4. Luật phân bố chuẩn $N(\mu;\sigma^2)$

a) Định nghĩa. BNN liên tục X nhận giá trị trên $\mathbb{R}$ được gọi là có phân bố chuẩn (hay phân bố theo luật chuẩn) với tham số μ vμ σ^2, ký hiệu là $X \sim N(\mu;\sigma^2)$, nếu hàm mật độ của nó có dạng

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}. \qquad (2.3.10)$$

Khảo sát hàm này ta nhận được các kết quả sau (xem Hình 2.9):



Hình 2.9. Mật độ của phân bố chuẩn $N(\mu;\sigma^2)$.

- Đồ thị nằm trên trục hoành.
- Trục Ox là tiệm cận ngang.
- Đồ thị đối xứng qua đường thẳng $x = \mu$; có dạng hình chuông.

- Cực đại $\dfrac{1}{\sqrt{2\pi\sigma^2}}$ đạt được tại duy nhất 1 điểm $x_0 = \mu$.
- Khi σ không đổi, μ tăng thì đồ thị tịnh tiến sang phải, μ giảm thì đồ thị tịnh tiến sang trái.
- Khi μ không đổi, σ tăng thì đồ thị thấp xuống, "tù ra"; σ giảm thì đồ thị cao lên, "nhọn hơn".

b) Các số đặc trưng

Định lý 2.8. Đối với $X \sim N(\mu;\sigma^2)$ th× $E[X]=\mu$; $V[X]=\sigma^2$.

Chứng minh. $E[X] = \int\limits_{-\infty}^{\infty} x \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\} dx.$

Đặt biến $z = \dfrac{x-\mu}{\sigma}$; $dx = \sigma dz$ ta được

$$E[X] = \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \sigma(\sigma z+\mu)\exp\left(-\frac{z^2}{2}\right)dz$$

$$= \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} z\exp\left(-\frac{z^2}{2}\right)dz + \mu\frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right)dz.$$

Số hạng đầu bằng không. Sử dụng tích phân Poisson $\int\limits_{-\infty}^{\infty} e^{-(z^2/2)}dz = \sqrt{2\pi}$, số hạng thứ hai bằng μ. Vậy $E[X] = \mu$.

Sử dụng tích phân từng phần và tương tự trên ta nhận được $V[X] = \sigma^2$. □

μ vµ σ^2 được gọi là 2 tham số đặc trưng của phân bố chuẩn.

c) BNN chuẩn hoá (chuẩn tắc).

BNN $Z \sim N(0;1)$ được gọi là chuẩn tắc, hàm mật độ cho bởi

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}. \qquad (2.3.12)$$

Hàm $\varphi(x)$ chẵn nên đồ thị của nó đối xứng qua trục tung. Giá trị của $\varphi(x)$ được lập bảng với $x \in [0;\ 4]$. Hàm phân bố tương ứng

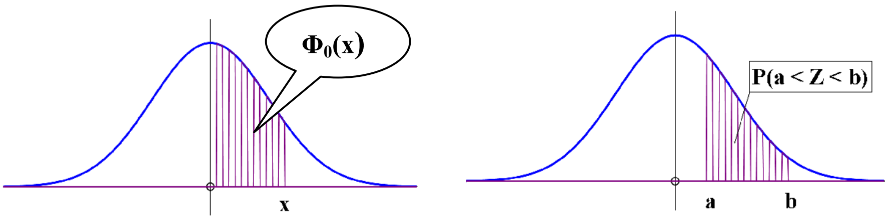
$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \qquad (2.3.13)$$

cũng được lập bảng. Tuy nhiên, để tiết kiệm bảng, thay cho $\Phi(x)$, người ta lập bảng các giá trị của hàm Laplace (Bảng A-1 ở Phụ lục)

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{x} e^{-\frac{t^2}{2}} dt \qquad (2.3.14)$$

với $x \in [0;\ 4]$. Với $x > 4$ coi $\Phi_0(x) \approx \dfrac{1}{2}$ (xem Hình 2.10). Rõ ràng

$\Phi_0(0) = 0;\ \Phi_0(-x) = -\Phi_0(x);$



Hình 2.10 Hàm Laplace và các xác suất chuẩn hóa

Khi cần tính F(x) qua $\Phi(x)$ hay ngược lại ta dùng công thức:

$$\Phi(x) = \frac{1}{2} + \Phi_0(x). \qquad (2.3.15)$$

Công thức sau có ích để tính xác suất BNN Z thuộc đoạn [a;b]:

$$P(a < Z < b) = \Phi_0(b) - \Phi_0(a). \qquad (2.3.16)$$

d) Biến đổi tuyến tính BNN chuẩn.

Cho $X \sim N(\mu;\sigma^2)$. Với $a, b \in \mathbb{R}$, $Y = aX + b$ có phân bố chuẩn. Từ đó dễ thấy $aX + b \sim N(a\mu + b;\ a^2\sigma^2)$ (xem Hệ quả của Định lý 3.12).

Hệ quả.

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0;\ 1). \qquad (2.3.17)$$

Hệ quả này cho ta phương pháp thuận lợi để tính $P\{X \in [a;b]\}$:

$$P\{X \in [a;b]\} = P\left\{\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right\} = \Phi_0(\frac{b-\mu}{\sigma}) - \Phi_0(\frac{a-\mu}{\sigma}) \qquad (2.3.18)$$

Ví dụ 2.15. Kích thước của các chi tiết do một máy sản xuất ra là BNN có phân bố chuẩn với trung bình 5cm và sai số (σ) là 0,9cm. Tìm xác suất để lấy ngẫu nhiên một chi tiết có kích thước nằm trong khoảng từ 4 đến 7cm.

Giải. Gọi X là kích thước chi tiết lấy ra, $X \sim N(5;\ 0{,}9^2)$. Ta có

$$P = P(4 < X < 7) = P\left(\frac{4-5}{0{,}9} < \frac{X-5}{0{,}9} < \frac{7-5}{0{,}9}\right) = \Phi_0(\frac{7-5}{0{,}9}) - \Phi_0(\frac{4-5}{0{,}9})$$

$$= \Phi_0(2{,}222) - \Phi_0(-1{,}111) = 0{,}4863 + 0{,}3665 = 0{,}8533.\ \#$$

e) Phân vị .

Định nghĩa. Cho BNN X có luật phân bố $\mathcal{L}$ nào đó. Phân vị mức α của X (hay của luật phân bố $\mathcal{L}$) là giá trị X_α xác định từ đẳng thức

$$P(X > X_\alpha) = \alpha.$$

Phân vị chuẩn mức α, ký hiệu z_α, là giá trị xác định bởi

$$P\{Z > z_\alpha\} = \alpha, \text{ với } Z \sim N(0;\ 1) \qquad (2.3.19)$$

hay
$$\frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{+\infty} e^{-\frac{t^2}{2}} dt = \alpha.$$

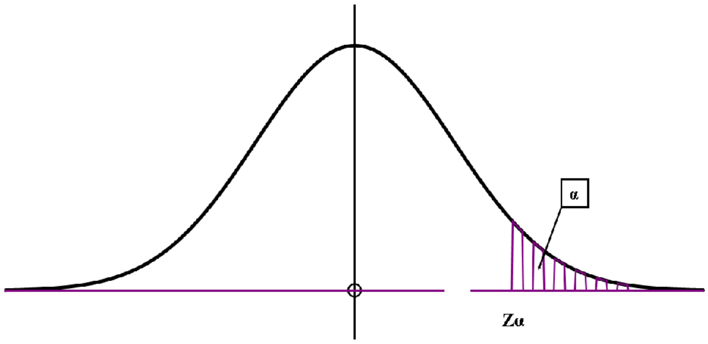
cũng vậy $\Phi_0(z_\alpha) = 0,5 - \alpha$.

Tính chất của phân vị chuẩn là

$$z_{1-\alpha} = -z_\alpha. \qquad (2.3.20)$$

Một số giá trị đặc biệt:

$$\begin{cases} z_{0,10} = 1,280; & z_{0,05} = 1,645; \\ z_{0,025} = 1,960; & z_{0,01} = 2,326. \end{cases} \qquad (2.3.21)$$



Hình 2.11. Phân vị chuẩn mức α

Lưu ý. Nhiều tài liệu không lập bảng của z_α mà lập bảng của p_α hoặc u_α với $P\{|Z| < p_\alpha\} = \alpha$; $P\{Z < u_\alpha\} = \alpha$; $Z \sim N(0;1)$.

f) Sai số trung gian, dạng mật độ dùng trong pháo binh.(Dành cho lớp Vũ khí, Đạn)

Dạng mật độ (2.3.23) của phân bố chuẩn hay được sử dụng trong Pháo binh.

g) Quy tắc 2σ (3σ)

Cho $X \sim N(\mu, \sigma^2)$; vì $\frac{X-\mu}{\sigma} \sim N(0,1)$, theo công thức (2.3.16) thì

$$P\{|X-\mu| < \varepsilon\} = P\left\{-\frac{\varepsilon}{\sigma} < \frac{X-\mu}{\sigma} < \frac{\varepsilon}{\sigma}\right\} = 2\Phi_0(\frac{\varepsilon}{\sigma}). (2.3.25)$$

Thay $\varepsilon = 1\sigma, 2\sigma, 3\sigma$ ta được

$$P\{|X-\mu| < 1\sigma\} = 2\Phi_0(1) = 0,68268;$$

$$P\{|X-\mu| < 2\sigma\} = 2\Phi_0(2) = 0,95450;$$

$$P\{|X-\mu| < 3\sigma\} = 2\Phi_0(3) = 0,9973. \qquad (2.3.26)$$

Các xác suất 0,9545; 0,9973 là rất lớn. Theo nguyên lý xác suất lớn ta có quy tắc 2σ (3σ) sau đây:

Quy tắc*:* Đối với BNN chuẩn thì hầu như chắc chắn (độ tin cậy trên 95% (trên 99%)), BNN chỉ sai lệch so với giá trị trung bình của nó một lượng không quá 2σ (3σ).

h) Tính phổ cập của phân bố chuẩn. Ta rất hay gặp phân bố chuẩn. Sở dĩ như vậy vì xảy ra Định lý giới hạn trung tâm (xem Định lý 3.25):

Nếu BNN X là kết quả của rất nhiều nguyên nhân, mỗi nguyên nhân chỉ có vai trò không đáng kể đến kết quả cuối cùng thì X có phân bố rất gần với phân bố chuẩn.

2.3.5. Luật phân bố mũ E(λ)

BNN X được gọi là có phân bố mũ với tham số λ, ký hiệu $X \sim E(\lambda)$, nếu hàm mật độ có dạng

$$f(x) = \begin{cases} 0 & \text{nếu } x < 0; \\ \lambda e^{-\lambda x} & \text{nếu } x \geq 0. \end{cases} \qquad (2.3.27)$$

Dễ chứng minh rằng

$$E[X] = \frac{1}{\lambda}; \quad V[X] = \frac{1}{\lambda^2}. \qquad (2.3.28)$$

Tuổi thọ của các thiết bị, tuổi thọ của nhiều loài sinh vật ... có phân bố mũ. Phân bố mũ có tính chất “không có trí nhớ” (xem Bài tập 2.15), nó còn được gọi là phân bố Markov.

2.3.6. Luật phân bố khi bình phương χ2(n)

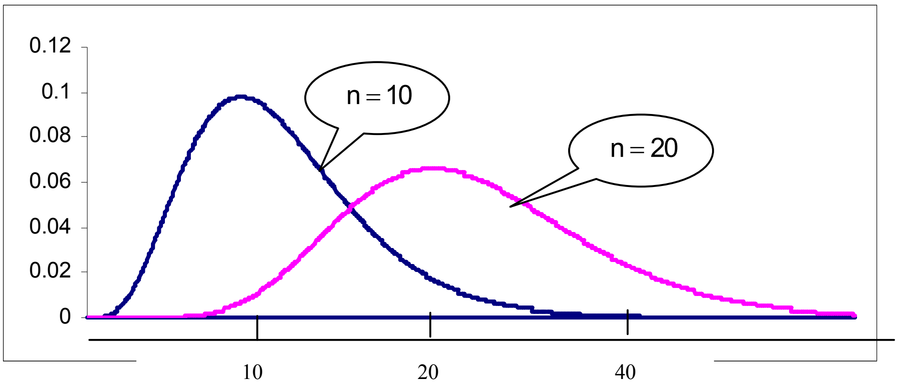
Định nghĩa. BNN liên tục X gọi là có phân bố χ^2 (khi bình phương) với n bậc tự do, ký hiệu $X \sim \chi^2(n)$, nếu hàm mật độ cho bởi (xem [1])… Ta có

$$\begin{cases} E[X] = n; \\ V[X] = 2n. \end{cases} \qquad (2.3.31)$$

Ngoài ra, với n lớn, phân bố χ^2 gần với phân bố chuẩn.

Nhận xét. Với n > 30, xấp xỉ phân bố $\chi^2(n)$ bởi phân bố chuẩn là tốt. Từ (2.3.31), phân bố chuẩn xấp xỉ là N(n; 2n). Hình 2.12 đưa ra hàm mật độ của phân bố χ^2 với $n = 10$ và $n = 20$ bậc tự do.

Nếu $X \sim \chi^2(n)$ thì $Y = aX$ $(a > 0)$ được gọi là có n bậc tự do.



Hình 2.12. Hàm mật độ của phân bô khi bình phương $\chi^2(n)$

. *Phân vị*. Phân vị mức α của phân bố khi bình phương với n bậc tự do, ký hiệu $\chi^2_\alpha(n)$, là giá trị xác định từ biểu thức

$$P(X > \chi^2_\alpha(n)) = \alpha, \quad (0 < \alpha < 1) \qquad (2.3.32)$$

trong đó $X \sim \chi^2(n)$. (Xem Hình 2.13(a)).

Giá trị $\chi^2_\alpha(n)$ có thể xem ở Bảng A-3, Phụ lục A.

2.3.7. Luật phân bố Student T(n)

Định nghĩa. BNN liên tục X gọi là có phân bố Student với n bậc tự do, ký hiệu $X \sim T(n)$, nếu hàm mật độ có dạng

Có thể tính được $E[T] = 0; \quad V[T] = n/(n-2), \; n = 3, 4, \ldots$

Phân vị. Phân vị Student mức α với n bậc tự do, ký hiệu $t_\alpha(n)$, là giá trị xác định từ biểu thức

$$P(X > t_\alpha(n)) = \alpha, \quad (0 < \alpha < 1). \qquad (2.3.34)$$

trong đó $X \sim T(n)$. (Xem Hình 2.13(b)).

Các giá trị $t_\alpha(n)$ đã được lập bảng với n và α khác nhau (xem Bảng A-2). Tính chất quan trọng của phân vị Student là

$$t_{1-\alpha}(n) = -t_\alpha(n); \quad t_\alpha(n) \approx z_\alpha \quad \text{ví i } n > 30.$$

2.3.8. Luật phân bố Fisher-Snedecor $F(n_1, n_2)$

Định nghĩa. BNN liên tục X gọi là có phân bố Fisher-Snedecor với n_1, n_2 bậc tự do, ký hiệu $X \sim F(n_1; n_2)$, nếu hàm mật độ có dạng … (2.3.35)

Phân vị. Phân vị mức α của phân bố F với n_1, n_2 bậc tự do, ký hiệu $f_\alpha(n_1; n_2)$ là giá trị xác định bởi

$$P(X > f_\alpha(n_1, n_2)) = \alpha, \quad \text{ví i } X \sim F(n_1; n_2). \qquad (2.3.36)$$

2.3.9. Các luật phân bố khác

Với luật phân bố rời rạc chúng ta hay gặp phân bố luỹ thừa, nhị thức âm, siêu hình học…. Với luật phân bố liên tục ta hay gặp phân bố tam giác, Weibull (1,2,3 tham số), Bê ta … Tóm tắt về các luật phân bố đưa ra ở Phụ lục B.

TÓM TẮT CHƯƠNG 2

BÀI TẬP: Các đặc trưng số của BNN (2 tiết)
Một số phân bố quan trọng (1 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 6: Vectơ ngẫu nhiên

Chương, mục: 3

Tiết thứ: 21-24 Tuần thứ: 6

Mục đích, yêu cầu:

- Tính được phân bố biên, phân bố điều kiện từ bảng Xs hay hàm mật độ đồng thời.
- Nắm được điều kiện cần đủ của 2 BNN độc lập thông qua bảng XS hay qua mật độ.

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

Bài tập về biến ngẫu nhiên

Chương 3. Véc tơ ngẫu nhiên

§ 3.1. Phân bố xác suất 2 chiều

§3.2. Phân bố điều kiện, sự độc lập của các BNN

BÀI TẬP: Một số phân bố quan trọng (1 tiết)

Chương 3
VÉC TƠ NGẪU NHIÊN

§3.1. PHÂN BỐ XÁC SUẤT HAI CHIỀU (1tiết)

3.1.1. Véc tơ ngẫu nhiên hai chiều

Định nghĩa. Véc tơ ngẫu nhiên (random vector) là véc tơ mà mỗi thành phần của nó là một BNN xác định trên cùng một không gian xác suất.

Trong thực tế người thường bỏ qua tính đo được của VTNN.

Việc xét VTNN là xét nhiều BNN đồng thời, không riêng rẽ. Chính vì thế, VTNN còn được gọi là BNN nhiều chiều.

3.1.2. Bảng xác suất của VTNN 2 chiều rời rạc

Bảng xác suất của VTNN 2 chiều rời rạc liệt kê các giá trị có thể của VTNN và xác suất tương ứng. Nếu

$$p(x_i, y_j) = p_{XY}(x_i, y_j) = P(X = x_i; Y = y_j),\ i, j = 1,2,...$$

thì bảng xác suất của (X,Y) thể hiện ở Bảng 3.1. Bảng này có thể hữu hạn hay vô hạn. Cột ngoài cùng ghi tổng các cột trước nó, dòng cuối ghi tổng các dòng trên nó. Các xác suất $p_{XY}(x_i, y_j)$ phải thoả mãn:

$$p_{XY}(x_i, y_j) \ge 0;\ \sum_{i,j} p_{XY}(x_i, y_j) = 1.$$

Từ bảng xác suất của (X,Y) ta có thể lập được bảng xác suất của các BNN thành phần. Thực vậy,

$$p_X(x_i) = P(X = x_i) = \sum_j P(X = x_i; Y = y_j) = \sum_j p_{XY}(x_i, y_j); \quad (3.1.3)$$

$$p_Y(y_j) = P(Y = y_j) = \sum_i P(X = x_i; Y = y_j) = \sum_i p_{XY}(x_i, y_j). \quad (3.1.4)$$

Bảng 3.1. Bảng xác suất của VTNN (X,Y)

X \ Y	y_1 . . . y_2 . . . . y_n .	$\sum$
x_1	$p(x_1,y_1)$ $p(x_1,y_2)$ . . . $p(x_1,y_m)$.	$p(x_1)$
x_2	$p(x_2,y_1)$ $p(x_2,y_2)$ . . . $p(x_2,y_m)$.	$p(x_2)$
.	. . . .	.
.	. . . .	.
x_n	$p(x_n,y_1)$ $p(x_n,y_2)$ . . . $p(x_n,y_m)$.	$p(x_n)$
.	. . . .	.
$\sum$	$p(y_1)$ $p(y_2)$ . . . $p(y_n)$.	1,000

Khi không sợ lầm ta bỏ đi các chỉ số. Bảng xác suất của X và Y là

X	x_1 x_2 . . . x_n . .	Y	y_1 y_2 . . . y_n .
P	$p(x_1)$ $p(x_2)$. . . $p(x_n)$. .	P	$p(y_1)$ $p(y_2)$. . . $p(y_n)$.

Chúng còn được gọi là các bảng xác suất biên. Rõ ràng

$$\sum_i p(x_i) = \sum_j p(y_j) = 1.$$

3.1.3. Hàm phân bố của VTNN 2 chiều

Hàm phân bố của VTNN (X,Y) (còn gọi là hàm phân bố đồng thời của các BNN X,Y), ký hiệu là $F_{XY}(x,y)$, được xác định bởi

$$F_{XY}(x,y) = P(X \le x, Y \le y), \quad x, y \in \mathbb{R}. \qquad (3.1.5)$$

Tính chất của hàm phân bố của VTNN khá giống với trường hợp 1 chiều.

Định lý 3.1. Giả sử $F_{XY}(x,y)$, $F_X(x)$, $F_Y(y)$ lần lượt là hàm phân bố của VTNN (X,Y), BNN X và BNN Y. Khi đó

i) $0 \le F_{XY}(x,y) \le 1, \quad x, y \in \mathbb{R}.$ (3.1.6)

ii) $F_{XY}(x,y)$ là hàm không giảm theo cả hai đối số:

$F_{XY}(x_1,y_1) \le F_{XY}(x_2,y_2)$ ví i $x_1 \le x_2,\ y_1 \le y_2$.(3.1.7)

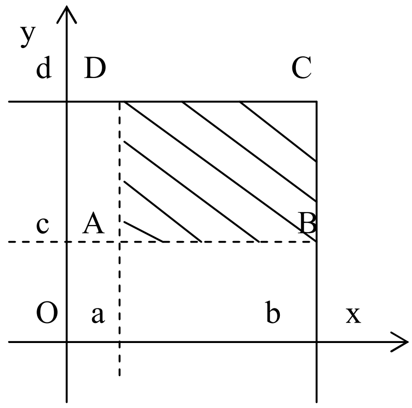
iv) Tại vô hạn thì

(xem [1])

v) Xác suất để (X,Y) rơi vào hình chữ nhật $\{(x,y): a < x \le b;\ c < y \le d\}$:

$$P(a < X \le b;\ c < Y \le d) = F(a,c) - F(a,d) - F(b,c) + F(b,d). \quad (3.1.10)$$

Các hàm phân bố $F_X(x)$, $F_Y(y)$ tính theo (3.1.9) gọi là phân bố biên của hàm phân bố $F_{XY}(x,y)$.



Hình 3.1. Xác suất để điểm ngẫu nhiên (X,Y) rơi vào hình chữ nhật

3.1.4. Hàm mật độ

Định nghĩa. Đạo hàm riêng hỗn hợp cấp hai của hàm phân bố, ký hiệu là $f_{XY}(x,y)$, được gọi là hàm mật độ của VTNN (X,Y):

$$f_{XY}(x,y)=\frac{\partial F_{XY}(x,y)}{\partial x\,\partial y}. \qquad (3.3.11)$$

Nhận xét. Nếu đạo hàm riêng ở vế phải của (3.1.11) tồn tại khắp nơi, có thể trừ ra trên một số hữu hạn đường cong thì ta có thể bổ sung giá trị tuỳ ý cho đạo hàm riêng trên các đường cong đó; coi mật độ xác định khắp nơi; VTNN được gọi là liên tục. Nếu đạo hàm riêng không tồn tại, có thể dùng hàm delta để xem xét khi tính toán các đạo hàm này.

Khi không sợ hiểu nhầm, ta có thể bỏ đi các chỉ số ở hàm phân bố cũng như ở hàm mật độ: Thay cho $F_{XY}(x,y)$, $f_{XY}(x,y)$... ta viết $F(x,y)$, $f(x,y)$ ...

Định lý 3.2. Hàm mật độ $f(x,y)$ của VTNN (X,Y) có tính chất sau

i) $$f(x,y)\ge 0,\quad x,y\in\mathbb{R}. \qquad (3.1.12)$$

ii) $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy=1. \qquad (3.1.13)$$

iii) Mật độ biên có thể tính thông qua hàm mật độ đồng thời:

$$f_X(x)=\int_{-\infty}^{\infty} f(x,y)dy;$$

$$f_Y(y)=\int_{-\infty}^{\infty} f(x,y)dx. \qquad (3.1.14)$$

iv) Hàm phân bố được tính từ hàm mật độ theo công thức

$$F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)dudv. \qquad (3.1.15)$$

v) Nếu B là một tập (đo được) bất kỳ trong $\mathbb{R}^2$ thì

$$P\big((X,Y)\in B\big)=\iint_B f(x,y)dxdy. \qquad (3.1.16)$$

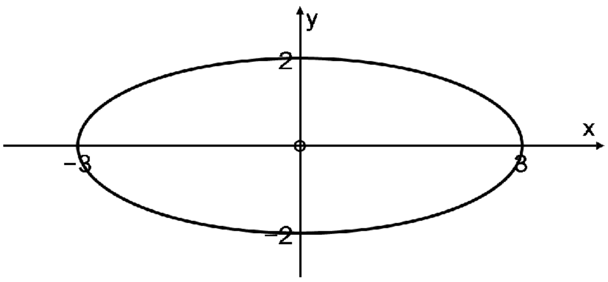
Ví dụ 3.3. Giả sử VTNN (X,Y) phân bố đều trên elíp $\frac{x^2}{9}+\frac{y^2}{4}\le 1$. Tìm các mật độ biên.

Giải. $f(x,y)=\begin{cases}\frac{1}{6\pi} & \text{nếu } \frac{x^2}{9}+\frac{y^2}{4}\le 1;\\ 0 & \text{trái lại.}\end{cases}$

$$|x|\le 3:\quad f_X(x)=\int_{-\infty}^{\infty} f(x,y)dxdy=\frac{1}{6\pi}\int_{-2\sqrt{1-x^2/9}}^{2\sqrt{1-x^2/9}} 1dx=\frac{2}{9\pi}\sqrt{9-x^2}.$$

$$\Rightarrow f_X(x)=\begin{cases}\frac{2}{9\pi}\sqrt{9-x^2} & \text{với } |x|\le 3\\ 0 & |x|>3\end{cases}.$$

$$f_Y(y)=\begin{cases}\frac{1}{2\pi}\sqrt{4-y^2} & \text{với } |y|\le 2\\ 0 & |y|>2.\end{cases}$$



§3.2. PHÂN BỐ ĐIỀU KIỆN, SỰ ĐỘC LẬP CỦA CÁC BNN (1 tiết)

3.2.1. Bảng phân bố điều kiện của BNN rời rạc

Định nghĩa. Giả sử VTNN rời rạc (X,Y) với $p(x_i,y_j)=P(X=x_i;Y=y_j)$, $i,j=1,2,\ldots$ Đặt

$$p(x_i|y_j)=P(X=x_i|Y=y_j)=\frac{P(X=x_i;Y=y_j)}{P(Y=y_j)}=\frac{p(x_i,y_j)}{p(y_j)}.\quad (3.2.1)$$

Cho i thay đổi ta thu được bảng xác suất điều kiện của X khi $Y=y_j$:

$X\|Y=y_j$	x_1	x_2	. . .	x_n	. .
P	$p(x_1\|y_j)$	$p(x_2\|y_j)$	. . .	$p(x_n\|y_j)$	. .

Nhận xét. Đây quả thực là bảng xác suất vì

$$0\le p(x_i|y_j)\le 1;\quad \sum_i p(x_i|y_j)=1.$$

Tương tự, ta có bảng xác suất điều kiện của Y khi $X=x_i$:

$Y\|X=x_i$	y_1	y_2	. . .	y_m	. .
P	$p(y_1\|x_i)$	$p(y_2\|x_i)$	. . .	$p(y_m\|x_i)$	. .

trong đó $p(y_j|x_i) = \dfrac{p(x_i, y_j)}{p(x_i)}$.

Ví dụ 3.4. Giả sử VTNN rời rạc (X,Y) có bảng xác suất

X \ Y	1	3	5
3	0,10	0,30	0,20
8	0,06	0,13	0,21

Dễ tìm được bảng xác suất điều kiện của Y khi $X = 3$ là

$Y\|X = 3$	1	3	5
P	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$

Nhận xét. Tập giá trị của BNN không thay đổi. Để tính xác suất tương ứng chỉ việc lấy dòng (cột) ở bảng xác suất đồng thời chia cho tổng dòng (tổng cột) đó.

3.2.2. Hàm phân bố điều kiện

b) Trường hợp Y liên tục.

Để đơn giản, ta coi cả X và Y liên tục với hàm mật độ đồng thời $f_{XY}(x, y)$.

Định nghĩa.

3.2.3. Sự độc lập của các BNN

Chúng ta nhớ lại rằng 2 biến cố A và B là độc lập nếu $P(AB) = P(A).P(B)$. Mở rộng sang BNN, ta coi 2 BNN là độc lập nếu mọi cặp biến cố liên kết với 2 BNN này là độc lập; cụ thể ta đưa ra định nghĩa sau đây.

Định nghĩa. Hai BNN X và Y được gọi là độc lập nếu với 2 tập (đo được) A, B bất kỳ trong $\mathbb{R}$ thì $(X \in A)$ vµ $(Y \in B)$ là 2 biến cố độc lập, tức là

$$P(X \in A; Y \in B) = P(X \in A).P(Y \in B). \qquad (3.2.10)$$

Việc kiểm tra đẳng thức (3.2.10) với mọi cặp A, B là khó khăn, thậm chí là không thể. Chúng ta đưa ra các tiêu chuẩn sau đây để kiểm tra tính độc lập rất thuận lợi trong thực hành.

Định lý 3.4. Hai BNN X và Y là độc lập khi và chỉ khi

$$F_{XY}(x, y) = F_X(x).F_Y(y), \quad x, y \in \mathbb{R}. \qquad (3.2.11)$$

Định lý 3.5. Các BNN X và Y độc lập khi và chỉ khi

$$f_{XY}(x, y) = f_X(x)f_Y(y), \quad x, y \in \mathbb{R}. \qquad (3.2.12)$$

Đặc biệt, nếu X và Y là hai BNN rời rạc với

$$p(x_i, y_j) = P(X = x_i; Y = y_j);\ p(x_i) = P(X = x_i);\ p(y_j) = P(Y = y_j)$$

thì X và Y độc lập khi và chỉ khi

$$p(x_i, y_j) = p(x_i).p(y_j), \quad i, j = 1, 2, \ldots \qquad (3.2.13)$$

Định lý 3.6. Nếu X và Y là 2 BNN độc lập còn g(x) và h(x) là 2 hàm (đo được) bất kỳ thì $U = g(X)$, $V = h(Y)$ là 2 BNN độc lập.

Ví dụ 3.6 *(Các biến ngẫu nhiên độc lập)*

Ví dụ 3.7 *(Các BNN không độc lập).* Xét 2 BNN X và Y với

3.2.4. Kỳ vọng điều kiện (xem [1])

*** BÀI TẬP: Một số phân bố quan trọng (tiếp 1 tiết)**

Phân bố XS 2 chiều (1 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	**Bài tập về nhà cho cả chương 3** Tài liệu [1]: 3(1 – <u>3</u> – 4 – <u>6</u> – <u>8</u> – 9 -<u>10</u>- <u>11</u> – 21 – <u>22</u> – 24 - 26- <u>27</u> – <u>33</u> – <u>38-</u> 40- 49- <u>53</u> - 54- 55). Tài liệu [2]: Tr 110-112: 10, 11, 14, 15, 16
Tài liệu	Tài liệu [1], tr ....

Bài giảng 7: Vectơ ngẫu nhiên (tiếp)

Chương, mục: 3

Tiết thứ: 25-28 Tuần thứ: 7

Mục đích, yêu cầu:

- Tính được kỳ vọng, phương sai, hiệp phương sai, hệ số tương quan.
- Ma trận hiệp phương sai và MT hệ số tương quan.
- Tìm được phân bố của 1 vài hàm đơn giản của VTNN.

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§3.3. Các số đặc trưng của VTNN

§3.5. Hàm của các BNN

§3.3. CÁC SỐ ĐẶC TRƯNG CỦA VTNN (1 tiết)

3.3.1. Các số đặc trưng của BNN thành phần

Nếu biết phân bố đồng thời của các BNN, ta có thể biết phân bố biên và từ đó ta có thể biết được kỳ vọng, phương sai của mỗi BNN đã cho. Tuy nhiên có thể tính trực tiếp các đặc trưng này từ phân bố đồng thời.

a) Trường hợp rời rạc.

$$E[X]=\sum_{i=1}^{\infty}x_i p(x_i)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}x_i p(x_i,y_j)$$

$$E[Y]=\sum_{j=1}^{\infty}y_j p(y_j)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}y_j p(x_i,y_j). \qquad (3.3.1)$$

$$V[X]=E[X^2]-(E[X])^2=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}x_i^2 p(x_i,y_j)-(E[X])^2;$$

$$V[Y]=E[Y^2]-(E[Y])^2=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}y_j^2 p(x_i,y_j)-(E[Y])^2. \qquad (3.3.2)$$

b. Trường hợp liên tục

$$E[X]=\int_{-\infty}^{\infty}xf_X(x)dx=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dxdy;$$

$$E[Y]=\int_{-\infty}^{\infty}yf_Y(y)dy=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf(x,y)dxdy. \qquad (3.3.3)$$

$$V[X]=E[(X-E[X])^2]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-E[X])^2 f(x,y)dxdy$$

$$= E[X^2] - (E[X])^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f(x,y)dxdy - (E[X])^2;$$

$$V[Y] = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y - E[Y])^2 f(x,y)dxdy$$

$$= E[Y^2] - (E[Y])^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f(x,y)dxdy - (E[X])^2. \quad (3.3.4)$$

3.3.2. Hiệp phương sai và hệ số tương quan (RẤT QUAN TRỌNG!)

Định lý 3.8. Giả sử (X,Y) là VTNN với hàm mật độ f(x,y) còn $\varphi(x,y)$ là hàm 2 biến sao cho $Z = \varphi(X,Y)$ là BNN có kỳ vọng. Khi đó

$$E[\varphi(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \varphi(x,y)f(x,y)dxdy. \quad (3.3.5)$$

Đặc biệt, nếu X và Y rời rạc với $p(x_i, y_j) = P(X = x_i, Y = y_j)$ thì

$$E[\varphi(X,Y)] = \sum_i \sum_j \varphi(x_i, y_j)p(x_i, y_j). \quad (3.3.6)$$

Mô men cấp (1,1) của X và Y,

$$m_{11} = E[XY]$$

được gọi là mô men tương quan gốc, gọi tắt là tương quan (correlation), của X và Y, ký hiệu bởi R[X,Y].

Hiệp phương sai (covariance) của X và Y, ký hiệu bởi Cov(X,Y):

$$Cov(X,Y) = E\{(X - E[X])(Y - E[Y])\}. \quad (3.3.8)$$

Khai triển vế phải ta thu được

$$E\{(X - E[X])(Y - E[Y])\}$$

$$= E[X^2] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \quad = E[XY] - E[X]E[Y].$$

Từ đó nhận được công thức sau rất tiện lợi để tính hiệp phương sai

$$Cov(X,Y) = E[XY] - E[X]E[Y]. \quad (3.3.9)$$

Hệ số tương quan (correlation coefficient) của X và Y, ký hiệu bởi $\rho(X,Y)$ (hay ρ_{XY}):

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V[X]}\sqrt{V[Y]}} = \frac{E(X - E[X])(Y - E[Y])}{\sqrt{V[X]}\,\sqrt{V[Y]}}. \quad (3.3.10)$$

Tính chất.

a) $\quad -1 \le \rho(X,Y) \le 1. \quad (3.3.11)$

b) Nếu $\rho(X,Y) = \pm 1$ thì BNN này biểu diễn tuyến tính qua BNN kia, cụ thể là, tồn tại 2 số thực a, b, a cùng dấu với $\rho(X,Y)$ để $Y = aX + b$.

Ta nói rằng X và Y **không tương quan** nếu $\rho(X,Y) = 0$. Điều này tương đương với

$$Cov(X,Y) = 0 \text{ hay } E[XY] = E[X]E[Y]. \quad (3.3.12)$$

Trái lại, X và Y gọi là tương quan với nhau.

Lưu ý: Nếu X và Y độc lập thì $E[XY] = E[X]E[Y]$ (xem (2.2.6)): Chúng không tương quan.

Điều ngược lại không đúng: Có những ví dụ chứng tỏ rằng, $Cov(X,Y) = 0$ nhưng X và Y không độc lập.

* Các ma trận

$$\begin{pmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{pmatrix}$$ **Ma trận hiêp phương sai**

$$\begin{pmatrix} \rho(X,X) & \rho(X,Y) \\ \rho(Y,X) & \rho(Y,Y) \end{pmatrix}$$ **Ma trận hệ số tương quan**

Có thể viết chúng dưới dạng

$$\begin{pmatrix} V[X] & Cov(X,Y) \\ Cov(X,Y) & V[Y] \end{pmatrix};$$

$$\begin{pmatrix} 1 & \rho(X,Y) \\ \rho(X,Y) & 1 \end{pmatrix}:$$ đối xứng, xác định không âm.

§3.5. HÀM CỦA CÁC BIẾN NGẪU NHIÊN (1 tiết)

3.5.1. Hàm vô hướng của một BNN

Cho X là BNN trên không gian xác suất (S, $\mathcal{F}$, P) còn $y = g(x)$ là hàm số (đo được) trên $\mathbb{R}$. Khi đó

$$Y = g(X)$$

là một BNN. Chúng ta muốn tìm luật phân bố của BNN Y này.

a) Trường hợp rời rạc.

Giả sử X là BNN rời rạc với bảng xác suất

X	x_1 . . . x_n . .
P	p_1 . . . p_n . .

Nếu $g(x_i) \neq g(x_j)$ với mọi $i \neq j$ thì có ngay bảng xác suất của Y

Y	$g(x_1)$ . . . $g(x_n)$ . .
P	p_1 . . . p_n . .

Nếu có một hoặc một số cột cột i, j mà $g(x_i) = g(x_j)$ thì ta chỉ việc gộp các cột này với xác suất bằng tổng các xác suất tương ứng.

Ví dụ 3.8. Cho bảng

X	0	1	2
P	0,2	0,3	0,5

Với $Y = X^2 - 2X - 3$ ta có

$P(Y = -4) = P(X^2 - 2X - 3 = -4) = P(X = 1) = 0,3;$

$P(Y = -3) = P(X^2 - 2X - 3 = -3) = P(X = 0) + P(X = 2) = 0,7.$

Y	-4	-3

P	0,3	0,7

Vậy bảng xác suất của Y:

Ta có thể thu được bảng này bằng cách lập bảng sau rồi gộp cột 2 với cột 4

X	0	1	2
Y	- 3	- 4	- 3
P	0,2	0,3	0,5

b) Trường hợp liên tục. Thường chúng ta phải tìm hàm phân bố của g(X) rồi đạo hàm hàm này để thu được hàm mật độ. Xét ví dụ sau.

Ví dụ 3.9. Phép biến đổi tuyến tính. Cho $Y = aX + b, (a \neq 0)$.

Với $a > 0$ thì

$$F(y) = P(aX + b \le y) = P(X \le (y-b)/a) = F_X((y-b)/a);$$

$$f_Y(y) = (F_Y(y))' = \frac{1}{a} f\left(\frac{y-b}{a}\right) = \frac{1}{|a|} f\left(\frac{y-b}{a}\right). \qquad (3.5.1)$$

Xét trường hợp $a < 0$, công thức (3.5.1) cũng đúng. #

Trong trường hợp tổng quát chúng ta có

$$F_Y(y) = P(g(X) \le y) = P\{X \in g^{-1}((-\infty; y])\}, \qquad (3.5.2)$$

trong đó $g^{-1}((-\infty; y])$ là nghịch ảnh của tập $(-\infty; y]$ qua ánh xạ g(.):

$$g^{-1}((-\infty; y]) = \{x \in \mathbb{R} : g(x) \le y\}. \qquad (3.5.3)$$

Theo (2.1.18), nếu X là liên tục với hàm mật độ $f_X(x)$ thì

$$F_{g(X)}(y) = \int\limits_{\{x:\, g(x) \le y\}} f_X(x)dx \qquad (3.5.4)$$

b1. g(x) là hàm đơn điệu.

Định lý 3.12. Giả sử X là BNN với tập giá trị là miền I và hàm mật độ $f_X(x)$ còn $g(x)$ là hàm đơn điệu, khả vi liên tục trên I sao cho tồn tại hàm ngược $x = g^{-1}(y)$ cũng khả vi liên tục. Khi đó BNN $Y = g(X)$ có hàm mật độ là

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d(g^{-1}(y))}{dy} \right|. \qquad (3.5.5)$$

Hệ quả. Nếu X là BNN chuẩn $N(\mu; \sigma^2)$ thì biến đổi tuyến tính $Y = aX + b$ của nó cũng là BNN chuẩn và $Y \sim N(a\mu + b; a^2\sigma^2)$.

Chứng minh. $f_X(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$. Theo (3.5.5),

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) = \ldots = \frac{1}{\sqrt{2\pi a^2\sigma^2}} \exp\left\{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}\right\}. \square$$

b2. g(x) không là hàm đơn điệu.

Nếu g(x) là hàm không đơn điệu thực sự nhưng ta có thể tách tập xác định của nó thành một số khoảng rời nhau:

$$I = I_1 \cup I_2 \cup ... \cup I_n, \qquad I_i \cap I_j = \varnothing \ (i \neq j),$$

sao cho trên mỗi khoảng I_i, g(x) là hàm đơn điệu thì ta vẫn có thể tìm được mật độ $f_Y(y)$ của BNN Y. Xét các ví dụ sau.

Ví dụ 3.10 *(Phép biến đổi cosin của phân bố đều).* Xét

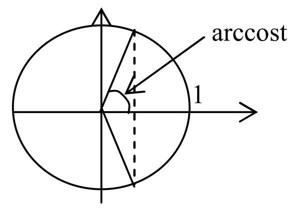
$$Y = \cos(X), \ \ X \sim U[0;2\pi).$$

Đây là trường hợp thường thấy trong hệ thống thông tin.

* $t \le -1$: $F_Y(t) = P(\cos(X) \le t) = 0 \Rightarrow f_Y(t) = 0$

* $t \ge 1$: $F_Y(t) = P(\cos(X) \le t) = 1.$

* $-1 < t < 1$: $F_Y(t) = P(\cos(X) \le t) =$



$$= P(\cos(X) \le t;\ 0 \le X < \pi)$$

$$+ P(\cos(X) \le t;\ \pi \le X < 2\pi)$$

$$= P(\arccos t \le X < \pi) + P(\pi \le X < 2\pi - \arccos t)$$

$$= \frac{1}{2\pi}[(\pi - \arccos t) + (2\pi - \arccos t - \pi)]$$

$$= \frac{-\arccos t}{\pi} \Rightarrow f_Y(t) = \frac{1}{\pi\sqrt{1-t^2}}. \qquad \#$$

Ví dụ 3.11 *(Phép biến đổi hớt ngọn* (clipping transformation)).

Định lý 3.13. Giả sử rằng X là BNN với hàm mật độ $f_X(x)$ còn hàm số $y = g(x)$ khả vi liên tục sao cho phương trình ẩn x:

$$y - g(x) = 0$$

có thể giải ra n nghiệm $x_k = x_k(y),\ k = 1,2,...,n$, các hàm $x_k(y)$ khả vi liên tục. Khi đó BNN $Y = g(X)$ có hàm mật độ là

$$f_Y(y) = \sum_{k=1}^{n} f_X(x_k(y)) \left| \frac{dx_k(y)}{dy} \right|. \qquad (3.5.6)$$

(Ở những nơi mà không có nghiệm thứ k thì lấy $f_X(x_k(y)) = 0$).

3.5.2. Hàm vô hướng của hai biến ngẫu nhiên

Định lý 3.14. Giả sử $z = g(x,y)$ là hàm 2 biến (đo được) còn $f_{XY}(x,y)$ là hàm mật độ đồng thời của 2 BNN X, Y. Khi đó, hàm phân bố của $Z = g(X,Y)$ là

$$F_Z(z) = \iint\limits_{\{(x,y):\ g(x,y) \le z\}} f_{XY}(x,y)dxdy. \qquad (3.5.7)$$

Nhận xét. i) Đạo hàm hai vế (3.5.7) ta nhận được mật độ $f_Z(z)$.

ii) (3.5.7) đúng với (X,Y) dạng bất kỳ (liên tục, rời rạc, hỗn hợp). Nếu X,Y rời rạc hay hỗn hợp, $f_{XY}(x,y)$ sẽ chứa hàm delta.

Ví dụ 3.12 *(Tổng hai biến ngẫu nhiên).* Xét trường hợp $Z = X + Y$. Thế thì $z = x + y$; công thức (3.5.7) trở thành

$$F_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_{XY}(x,y)dxdy.$$

Đạo hàm 2 vế theo z ta nhận được

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(z-y,y)dy. \qquad (3.5.8)\#$$

Hệ quả. Nếu hai BNN X và Y độc lập với mật độ tương ứng là $f_X(x)$ vµ $f_Y(y)$ thì tổng $Z = X + Y$ có mật độ là

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(x-t)f_Y(t)dt$$

$$= \int_{-\infty}^{\infty} f_Y(x-t)f_X(t)dt. \qquad (3.5.9)$$

Chứng minh. Vì $f_{XY}(x,y) = f_X(x)f_Y(y)$, áp dụng (3.5.8) thì

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_{XY}(z-y,y)dy = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy,$$

ta nhận được đẳng thức thứ nhất của (3.5.9). Đẳng thức thứ hai suy từ tính đối xứng. □

Tích phân ở vế phải của (3.5.9) có ý nghĩa quan trọng, thường xuyên được sử dụng sau này, gọi là **tích chập** của 2 hàm $f_X(x)$ vµ $f_Y(y)$. Tích chập được định nghĩa không chỉ cho hai hàm mật độ mà còn cho hai hàm bất kỳ.

Định nghĩa. Tích chập (tên khác: tích phân chập) của hai hàm f(x) và g(x), ký hiệu $f(x)*g(x)$ là hàm số xác định bởi

$$f(x)*g(x) = \int_{-\infty}^{\infty} f(x-t)g(t)dt \qquad (3.5.10)$$

(nếu tích phân ở vế phải tồn tại).

Ngoài tính chất tuyến tính, tích chập còn có tính chất giao hoán:

$$f(x)*g(x) = g(x)*f(x).$$

Như vậy, hệ quả trên khẳng định rằng, mật độ của tổng 2 BNN độc lập bằng tích chập của 2 mật độ thành phần:

$$f_{X+Y}(x) = f_X(x)*f_Y(x). \qquad (3.5.11)$$

3.5.3. Véc tơ hàm của véc tơ ngẫu nhiên

Giả sử (X,Y) là VTNN với luật phân bố đã biết, ta đi tìm phân bố đồng thời của $U = u(X,Y)$, $V = v(X,Y)$ với $u = u(x,y)$, $v = v(x,y)$ là 2 hàm cho trước.

Định lý 3.15. Giả sử rằng VTNN (X,Y) có hàm mật độ là $f_{XY}(x,y)$ và ánh xạ $(x,y) \mapsto (u,v)$ xác định trên tập giá trị của VTNN (X,Y) cho bởi

$$\begin{cases} u = u(x,y) \\ v = v(x,y) \end{cases} \qquad (3.5.16)$$

là song ánh với ánh xạ ngược

$$\begin{cases} x = x(u,v) \\ y = y(u,v) \end{cases} \qquad (3.5.17)$$

sao cho hai hàm $x = x(u,v)$, $y = y(u,v)$ khả vi liên tục, khi đó

$$f_{UV}(u,v) = f_{XY}(x(u,v), y(u,v))\left|\frac{D(x,y)}{D(u,v)}\right|. \qquad (3.5.18)$$

trong đó $\dfrac{D(x,y)}{D(u,v)}$ là định thức Jacobi của phép biến đổi (3.5.17).

Khi (3.5.16) không là song ánh có thể sử dụng định lý sau.

Định lý 3.16. Giả sử rằng

i) VTNN (X,Y) có hàm mật độ $f_{XY}(x,y)$.

ii) Các hàm số $u = u(x,y)$, $v = v(x,y)$ xác định trên tập giá trị Ω của VTNN (X,Y) sao cho với mọi $(x,y) \in \Omega$, hệ phương trình

$$\begin{cases} u - u(x,y) = 0 \\ v - v(x,y) = 0 \end{cases} \qquad (3.5.19)$$

có thể giải ra k nghiệm

$$\begin{cases} x_i = x_i(u,v) \\ y_i = y_i(u,v) \end{cases} \quad i = 1,...,k$$

trong đó các hàm $x_i(u,v)$; $y_i(u,v)$ khả vi liên tục với định thức Jacobi

$$\frac{D(x_i,y_i)}{D(u,v)} = \det\begin{pmatrix} \dfrac{\partial x_i}{\partial u} & \dfrac{\partial x_i}{\partial v} \\ \dfrac{\partial y_i}{\partial u} & \dfrac{\partial y_i}{\partial v} \end{pmatrix}.$$

Khi đó, (U,V) là VTNN với hàm mật độ

$$f_{UV}(u,v) = \sum_{i=1}^{k} f_{XY}(x_i(u,v), y_i(u,v))\left|\frac{D(x_i,y_i)}{D(u,v)}\right|. \qquad (3.5.20)$$

(Nếu với (u,v) mà hệ (3.5.17) không có nghiệm thứ j: $x = x_j(u,v)$, $y = y_j(u,v)$ thì coi $f_{XY}(x_j(u,v), y_j(u,v)) = 0$).

§3.6. MỘT SỐ PHÂN BỐ NHIỀU CHIỀU

3.6.1. Luật phân bố Poisson nhiều chiều

3.6.2. Luật phân bố chuẩn nhiều chiều

BÀI TẬP: Phân bố xác suất 2 chiều (1 tiết)

Phân bố điều kiện (1 tiết)

b) Thảo luận	
c) Tự học	**§3.6. MỘT SỐ PHÂN BỐ NHIỀU CHIỀU**
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 8: Vectơ ngẫu nhiên (tiếp)

Chương, mục: 3

Tiết thứ: 29-32 Tuần thứ: 8

Mục đích, yêu cầu:

- Nắm được 4 dạng hội tụ và mối quan hệ giữa chúng
- Các luật số lớn và định lý giới hạn trung tâm
- Làm được bài tập

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§3.7. Sự hội tụ của dãy các biến ngẫu nhiên

Bài tập về vector ngẫu nhiên

§3.7. SỰ HỘI TỤ CỦA DÃY CÁC BIẾN NGẪU NHIÊN (1 tiết)

3.7.1. Các dạng hội tụ

a) Định nghĩa. Giả sử X và X_n, n = 1, 2, … là các BNN cùng xác định trên không gian xác suất (S, $\mathcal{F}$, P).

i) Ta nói dãy các BNN $\{X_n\}$ hội tụ chắc chắn tới BNN X nếu

$$\lim_{n\to\infty} X_n(\zeta) = X(\zeta), \quad \forall \zeta \in S.$$

ii) Ta nói dãy các BNN $\{X_n\}$ hội tụ hầu chắc chắn tới BNN X và viết $X_n \xrightarrow{a.s.} X$ (hay $X_n \to X$ (a.s.)) nếu tồn tại biến cố $A \subset S$ với P(A) = 1 sao cho

$$\lim_{n\to\infty} X_n(\zeta) = X(\zeta), \quad \forall \zeta \in A$$

(a.s. là chữ viết tắt của almost sure).

iii) Ta nói dãy các BNN $\{X_n\}$ hội tụ theo xác suất tới BNN X và viết $X_n \xrightarrow{P} X$ nếu

$$\lim_{n\to\infty} P\{|X_n - X| \geq \varepsilon\} = 0, \quad \forall \varepsilon > 0.$$

iv) Ta nói dãy các BNN $\{X_n\}$ hội tụ trung bình cấp p $(0 < p < \infty)$ tới BNN X và viết $X_n \xrightarrow{L_p} X$ (hay $X_n \to X$ theo trung bình cấp p) nếu

$$E|X_n|^p < \infty, \ \forall n \quad \text{và} \quad \lim_{n\to\infty} E|X_n - X|^p = 0.$$

v) Ta nói dãy các BNN $\{X_n\}$ hội tụ theo luật (tên khác: theo phân bố) đến BNN X và ta viết $X_n \xrightarrow{L} X$ hay $F_{X_n} \rightrightarrows F_X$ nếu

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

tại mọi điểm liên tục của hàm phân bố $F_X(x)$.

Hội tụ trung bình cấp hai là quan trọng nhất. Hội tụ trung bình cấp hai còn gọi là hội tụ bình phương trung bình hay M.S. - hội tụ (mean square convergence), ký hiệu là

$$X_n \xrightarrow{\text{M.S.}} X, \ (n \to \infty) \qquad \text{hay } \text{l.i.m.}\, X_n = X$$

(l.i.m. là viết tắt của chữ limit in mean).

Riêng với hội tụ theo luật, các BNN X_n và X có thể xác định trên những không gian xác suất khác nhau.

c. Mối quan hệ giữa các dạng hội tụ

Định lý sau nêu lên mối quan hệ giữa các dạng hội tụ nêu trên.

Định lý 3.25. i) Dãy $\{X_n\}$ hội tụ hầu chắc chắn sẽ hội tụ theo xác suất:

$$X_n \xrightarrow{\text{a.s.}} X \ \Rightarrow \ X_n \xrightarrow{P} X.$$

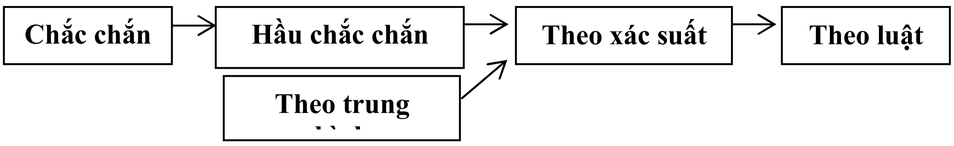
ii) Dãy $\{X_n\}$ hội tụ bình phương trung bình thì cũng hội tụ theo xác suất:

$$X_n \xrightarrow{\text{M.S.}} X \ \Rightarrow \ X_n \xrightarrow{P} X.$$

iii) Dãy $\{X_n\}$ hội tụ theo xác suất thì cũng hội tụ theo luật:

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X.$$

Mối quan hệ giữa các dạng hội tụ thể hiện ở giản đồ sau



3.7.2. Các định lý giới hạn

a. Bất đẳng thức Chebyshev

Giả sử X là BNN với kỳ vọng E[X] và phương sai V[X] hữu hạn. Khi đó, $\forall \varepsilon > 0$ xảy ra bất đẳng thức:

$$P\{|X - E[X]| \ge \varepsilon\} \le \frac{V[X]}{\varepsilon^2}. \tag{3.7.1}$$

Chứng minh. Ta chứng minh cho trường hợp X liên tục. Ta có:

$$V[X] = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x)dx \ge \int_{-\infty}^{E[X]-\varepsilon} (x - E[X])^2 f(x)dx$$

$$+ \int_{E[X]+\varepsilon}^{+\infty} (x - E[X])^2 f(x)dx \ge \varepsilon^2 P\{|X - E[X]| \ge \varepsilon\}.$$

Nhận được đpcm. □

Đôi khi dạng sau đây của (3.7.1) cũng rất tiện lợi:

$$P\{|X - E[X]| < \varepsilon\} \geq 1 - \frac{V[X]}{\varepsilon^2}.$$

Đặc biệt, khi $\varepsilon = n\sigma$ $(\sigma = \sqrt{V[X]})$ chúng ta thu được

$$P\{|X - E[X]| < n\sigma\} \geq 1 - \frac{1}{n^2}.$$

Nếu chọn n = 3 thì

$$P\{|X - E[X]| < 3\sigma\} \geq \frac{8}{9}. \qquad (3.7.2)$$

Bất đẳng thức (3.7.2) cũng được phát biểu dưới dạng quy tắc 3σ:

Mỗi BNN không lệch khỏi giá trị trung bình của nó một lượng 3σ với xác xuất khá lớn.

Chúng ta thấy xác suất "khá lớn" ở đây chỉ là 8/9, thấp hơn rất nhiều so với 0.9973 ở trường hợp X có phân bố chuẩn theo công thức (2.3.26). Như vậy, nếu biết thêm thông tin về tính chuẩn của BNN X, *chúng ta có những khẳng định mạnh hơn về khả năng xuất hiện biến cố* $\{|X - E[X]| < 3\sigma\}$.

b. Luật yếu số lớn

Định lý 3.27. Cho dãy BNN $\{X_n\}$ độc lập, cùng phân bố với kỳ vọng $E[X_i] = m$ và phương sai $V[X_i] = \sigma^2$ hữu hạn. Khi đó, với mọi $\varepsilon > 0$ cố định thì

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1 + ... + X_n}{n} - \mu\right| < \varepsilon\right\} = 1. \qquad (3.7.3)$$

Chứng minh. Từ giả thiết suy ra

$$E\left[\frac{X_1 + ... + X_n}{n}\right] = \mu; \quad V\left[\frac{X_1 + ... + X_n}{n}\right] = \frac{\sigma^2}{n}.$$

Áp dụng bất đẳng thức Chebyshev chúng ta có

$$P\left\{\left|\frac{X_1 + ... + X_n}{n} - \mu\right| < \varepsilon\right\} \geq 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

Chuyển qua giới hạn khi $n \to \infty$ ta nhận được đpcm. □

Theo các dạng hội tụ xét đến ở mục 3.7.1, luật yếu số lớn chính là:

Đối với dãy BNN độc lập, cùng phân bố với phương sai hữu hạn, dãy trung bình cộng hội tụ theo xác suất đến kỳ vọng chung của dãy.

Định lý này được nêu ra bởi Bernoulli ở cuối thế kỷ 17 như là thành công đầu tiên của lý thuyết xác suất non trẻ. Thực ra, với cùng giả thiết, chúng ta còn thu được sự hội tụ hầu chắc chắn, dạng hội tụ mạnh hơn hội tụ theo xác suất. Đó là nội dung của luật mạnh số lớn, công trình thuộc về Kolmogorov.

c. Luật mạnh số lớn

Định lý 3.28. Cho dãy BNN $\{X_n\}$ độc lập, cùng phân bố với kỳ vọng $E[X_i] = \mu$ và phương sai $V[X_i] = \sigma^2$ hữu hạn. Khi đó

$$P\left\{\lim_{n\to\infty}\frac{X_1+...+X_n}{n}=\mu\right\}=1. \qquad (3.7.4)$$

Như vậy, với điều kiện nêu ra, dãy trung bình cộng $\left(X_1+...+X_n\right)/n$ hội tụ hầu chắc chắn đến kỳ vọng μ.

d) Định lý giới hạn trung tâm

Hệ quả *(Định lý giới hạn Moivre - Laplace)*

Giả sử Z_n là BNN có phân bố nhị thức B(n;p) (0 < p < 1). Đặt:

$$T_n=\frac{Z_n-np}{\sqrt{npq}}. \qquad (3.7.7)$$

Khi đó

$$\lim_{n\to\infty}P\{T_n<x\}=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt. \qquad (3.7.8)$$

Nói cách khác, $\{T_n\}$ hội tụ theo luật đến phân bố chuẩn tắc.

Như vậy, với n lớn chúng ta có xấp xỉ

$$P\{T_n<x\}\approx F(x) \qquad (3.7.9)$$

với F(x) là hàm phân bố chuẩn tắc.

Bây giờ ta quay trở lại tính xấp xỉ $P_n(k_1\div k_2)$ như đã đưa ra ở (1.3.8). Ta có

$$P_n(k_1\div k_2)=P\{k_1\le S_n\le k_2\}$$

$$=P\left\{\frac{k_1-np}{\sqrt{npq}}\le\frac{S_n-np}{\sqrt{npq}}\le\frac{k_2-np}{\sqrt{npq}}\right\}\approx\Phi(\frac{k_2-np}{\sqrt{npq}})-\Phi(\frac{k_1-np}{\sqrt{npq}}).$$

Cuối cùng, vì $\Phi(x)=1/2+\Phi_0(x)$ chúng ta nhận được xấp xỉ

$$P_n(k_1\div k_2)\approx\Phi_0(x_2)-\Phi_0(x_1), \qquad (3.7.10)$$

trong đó $x_i=\frac{k_i-np}{\sqrt{npq}},\quad i=1,2$

$$\Phi_0(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt \text{ - } \quad \text{hàm Laplace.}$$

Người ta thấy rằng xấp xỉ là tốt nếu np > 5; nq > 5 hoặc npq > 20.

Ví dụ 3.21. Xác suất trúng đích của một xạ thủ khi bắn một viên đạn vào bia là 0,75. Tính xác suất để xạ thủ đó bắn 100 viên có 81 phát trúng đích trở lên.

Giải. np, nq > 10, chúng ta có thể áp dụng các kết quả nêu trên.

$$P=P_{100}(81\div100)\approx\Phi_0(x_2)-\Phi_0(x_1);$$

$$x_1=\frac{81-100.75}{\sqrt{100.0,75.0.25}}\approx1,38;\ x_2=\frac{100-100.75}{\sqrt{100.0,75.0.25}}\approx5,77.$$

Tra bảng ta có

$$\Phi_0(1,38)=0,4162;\ \ \Phi_0(5,77)\approx\Phi_0(3,0)=0,5000;$$

$$\Rightarrow P = 0,5000 - 0,4162 = 0,0838 \approx 8\% .$$

Định lý 3.30 *(Định lý giới hạn Poisson).*

Tự đọc (xem [1])

Ví dụ 3.22. Xác suất để trong 1 lần bay, một loại máy bay trên một tuyến đường nhất định bị tai nạn là $p = 10^{-4}$. Tìm xác suất để trong 1000 lần bay có: a) một lần bị tai nạn; b) ít nhất 1 lần bị tai nạn.

Giải. Ta có thể áp dụng Định lý giới hạn Poisson với n = 1000, p = 0,0001; $\lambda = np = 0,1$.

$$P_a = P_{1000}(1) \approx e^{-0,1}\frac{(0,1)^1}{1!} = 0,0905 \approx 9\%.$$

$$P_b = P_{1000}(1 \div 1000) = 1 - P_{1000}(0) \approx 1 - e^{-0,1}\frac{(0,1)^0}{0!} = 0,0952 \approx 9,5\%.$$

Đây là những xác suất khá nhỏ.

Nếu bây giờ p = 0,001, tính toán tương tự ta được $P_a \approx 0,3679$; $P_b \approx 0,6321$. Đây lại là những xác suất khá lớn.

BÀI TẬP: Các số đặc trưng của VTNN(2 tiết)
Hàm của các BNN (1 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 9: Kiểm định giả thuyết và ước lượng tham số

Chương, mục: 4

Tiết thứ: 33-36 Tuần thứ: 9

Mục đích, yêu cầu:

- Một số hiểu biết về mẫu, các phép lấy mẫu tốt
- Các tiêu chuẩn ƯL: không chệch, vững, hiệu quả, hợp lý

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

Ôn tập phần xác suất

Chương 4. Kiểm định giả thuyết và ước lượng tham số

§ 4.1 Mẫu và các đặc trưng mẫu

§ 4.2. Ước lượng điểm

Ôn tập phần xác suất

Chương 4. KIỂM ĐỊNH GIẢ THUYẾT VÀ ƯỚC LƯỢNG THAM SỐ

§4.1. MẪU VÀ CÁC ĐẶC TRƯNG MẪU (1 tiết)

4.1.1. Mẫu ngẫu nhiên

a) Định nghĩa

Để nghiên cứu về BNN X, người ta quan sát nó n lần. Giả sử các giá trị quan sát được là $X_1,...,X_n$. Véc tơ ngẫu nhiên $(X_1,...,X_n)$ được gọi là một mẫu quan sát về BNN X.

Định nghĩa. Giả sử X là BNN với hàm phân bố F(x). Véc tơ ngẫu nhiên $(X_1,...,X_n)$ với các thành phần $X_1,...,X_n$ độc lập, cùng hàm phân bố F(x) được gọi là một mẫu ngẫu nhiên đơn giản về BNN X;

N: kích thước mẫu; X: BNN gốc.

Mở rộng sang trường hợp X là một VTNN.

b) Các phương pháp lấy mẫu

b1. Tiến hành các thí nghiệm vật lý độc lập

Giả sử cần tìm hiểu về đại lượng vật lý X nào đó, nhà nghiên cứu tiến hành n thí nghiệm để xác định nó một cách cẩn thận nhất có thể, các kết quả đo đạc là độc lập với nhau. Ký hiệu X_i là kết quả đo ở thí nghiệm thứ i. VTNN $(X_1,...,X_n)$ thu được là một mẫu ngẫu nhiên đơn giản về X.

Đây là phương pháp cơ bản trong kỹ thuật. Tuy nhiên nhiều khi ta không thể tiến hành thí nghiệm được, mà chỉ có thể tiến hành các quan sát. Khi ấy các phương pháp thu nhận mẫu sau đây thường được sử dụng.

b2. Chọn mẫu hoàn lại, không hoàn lại

Ta hình dung các phương pháp này bằng 2 ví dụ sau đây.

Ví dụ 4.1. Giả sử chúng ta có một lô hàng gồm N sản phẩm, trong đó có m chính phẩm và $N-m$ phế phẩm. Đối với mỗi sản phẩm trong lô hàng, để ý đến chất lượng của nó ta đặt

$$X=\begin{cases}1 & \text{nÕu s¶n phÈm lµ tèt;}\\ 0 & \text{nÕu s¶n phÈm lµ háng.}\end{cases}$$

Để nghiên cứu về X, ta rút ngẫu nhiên (sao cho mỗi cá thể trong tổng thể đều có khả năng tham gia vào mẫu với xác suất như nhau) lần đầu một sản phẩm, kiểm tra chất lượng của nó rồi đặt

$$X_1=\begin{cases}1 & \text{nÕu s¶n phÈm lµ tèt;}\\ 0 & \text{nÕu s¶n phÈm lµ háng.}\end{cases}$$

Sau đó trả lại sản phẩm vào lô hàng. Lặp lại n lần rút như trên, gọi kết quả thu được là $(X_1,...,X_n)$. Đây là mẫu ngẫu nhiên theo phép lấy mẫu **hoàn lại**. Đó là 1 mẫu ngẫu nhiên đơn giản về X. #

Ví dụ 4.2. Mẫu ngẫu nhiên hoàn lại có nhược điểm là một phần tử có thể xuất hiện một số lần trong mẫu. Khi tập nghiên cứu không thật đồng nhất và số phần tử N của nó khá lớn, người ta hay dùng phương pháp chọn mẫu không hoàn lại. Cùng với các giả thiết ở Ví dụ 4.1 nhưng sau mỗi lần rút sản phẩm, đem kiểm tra chất lượng, ghi nhận kết quả, ta **không trả** sản phẩm về lại lô hàng nữa mà rút tiếp sản phẩm khác. Kết quả thu được $(X_1,...,X_n)$ gọi là mẫu ngẫu nhiên về X theo phương pháp không hoàn lại. #

Tổng quát. Bản chất bấtkỳ;

Có nhiều dấu hiệu quan tâm: mẫu ngẫu nhiên về VTNN.

Nhận xét.

Có thể chứng minh được, nếu $n \ll N$ thì mẫu ngẫu nhiên theo phương pháp lấy mẫu không hoàn lại có các thành phần "gần như độc lập với nhau" và "gần như cùng phân bố" với X. Nếu

$$n \le N/10$$

thì trong ứng dụng có thể xem **hai phương pháp lấy mẫu cho kết quả tốt như nhau: Mẫu theo phương pháp không hoàn lại cũng có các thành phần xem là độc lập và** cùng phân bố với X.

Phương pháp chọn mẫu không hoàn lại tương đương với chọn ngẫu nhiên đồng thời n phần tử từ tổng thể.

Nếu ta đã có một bản danh sách về các đối tượng cần nghiên cứu, ta có thể chọn ngẫu nhiên n phần tử từ danh sách bằng cách sử dụng dãy số ngẫu nhiên (từ bảng các số ngẫu nhiên hay tạo ra từ máy tính bất kỳ).

Chẳng hạn, ta cần chọn ra 9 phần tử từ danh sách gồm 430 phần tử đã đánh số. Ta tạo ra một dãy số ngẫu nhiên, mỗi số gồm 3 chữ số. Ví dụ, kết quả là:

346, 958, 034, 804, 501, 523, 047, 088, 034, 800, 419, 648, 907,
861, 822, 909, 928, 356, 210, 850, 569, 098, 258, 934, 487, …

Ta hãy chọn lần lượt từ trái qua phải, số nào lớn hơn 430 hoặc đã có mặt ở trước đó thì bỏ qua, số nào nhỏ hơn hay bằng 430 thì đưa vào mẫu cho đến khi đủ 9 phần tử thì dừng lại. Các phần tử của mẫu sẽ là

346, 034, 047, 088, 419, 356, 210, 098, 258.

b3. Lập mẫu theo cách phân tầng.

b4. Lập mẫu theo hệ thống.

Còn cách chọn mẫu khác là chọn mẫu theo kinh nghiệm

Tuỳ từng trường hợp mà chúng ta lựa chọn phương pháp chọn mẫu phù hợp. Lựa chọn phương pháp thu nhận mẫu luôn được nhà nghiên cứu tính đến, song chúng ta phải nhớ rằng quan sát (ngụ ý rằng quan sát chính xác) mới là điều quan trọng nhất. Nhấn mạnh vai trò của quan sát, Desabie - một tác giả có tiếng trong lĩnh vực điều tra thăm dò - đã viết: "Một cuộc điều tra thăm dò thất bại thì thường là do các sai lệch trong quan sát chứ không phải do các sai lầm trong việc lập mẫu".

Từ nay, giả sử rằng chúng ta có mẫu ngẫu nhiên đơn giản về BNN X, và chỉ xét những mẫu như vậy. Để cho gọn, chúng ta dùng từ 'mẫu" để chỉ mẫu ngẫu nhiên đơn giản. Có thể mẫu chưa phải là mẫu đơn giản, khi đó nhà nghiên cứu phải điều chỉnh hoặc hiểu kết quả thu được một cách phù hợp.

Khi đã có các quan sát về BNN rồi, tức là đã có kết quả ζ của thí nghiệm ngẫu nhiên) thì mẫu sẽ là một bộ số $(X_1(\zeta),...,X_n(\zeta))$, đó là một bộ n số cụ thể nào đó. Ta gọi đây là mẫu cụ thể (hay mẫu quan sát), thường được ký hiệu bởi những chữ cái nhỏ tương ứng, chẳng hạn $(x_1,...,x_n)$ trong trường hợp này.

4.1.2. Một số đặc trưng mẫu

a) Hàm phân bố thực nghiệm

Định nghĩa. Hàm phân bố thực nghiệm (còn gọi là hàm phân bố mẫu) từ mẫu $(X_1,...,X_n)$ về BNN X là hàm số

$$F_n(x) = \begin{cases} 0 & \text{nếu } x < \min(X_1,...,X_n) \\ \dfrac{k}{n} & \text{nếu có k phần tử trong mẫu } \le x \\ 1 & \text{nếu } \max(X_1,...,X_n) \le x. \end{cases}$$

***Định lý 4.1** (Glivenko).* Nếu $(X_1,...,X_n)$ là một mẫu từ BNN có hàm phân bố F(x) thì hàm phân bố thực nghiệm $F_n(x)$ tương ứng hội tụ đều hầu chắc chắn về F(x), cụ thể là:

$$P\{\sup_{x\in\mathbb{R}}|F_n(x)-F(x)| \to 0,\ (n\to\infty)\} = 1. \qquad (4.1.1)$$

Ý nghĩa. Khi n lớn, ta có thể dùng xấp xỉ $F(x) \approx F_n(x),\ x \in \mathbb{R}$.

b) Kỳ vọng mẫu

Định nghĩa. Kỳ vọng mẫu (hay trung bình mẫu) của BNN X ứng với mẫu ngẫu nhiên $(X_1,...,X_n)$ ký hiệu bởi $\overline{X}$ và được tính theo công thức

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + ... + X_n). \qquad (4.1.3)$$

Tính chất. Nếu $E[X] = \mu,\ V[X] = \sigma^2$ thì

$$E[\overline{X}] = \mu;$$
$$V[\overline{X}] = \sigma^2 / n. \qquad (4.1.4)$$

Chứng minh. . □

Nhận xét. Cả hai BNN X và $\overline{X}$ đều tập trung quanh kỳ vọng μ, song kỳ vọng mẫu $\overline{X}$ tập trung quanh μ mạnh hơn, ít phân tán hơn so với X quanh μ. Vậy, ta dùng $\overline{X}$ làm xấp xỉ cho μ lợi hơn dùng X.

Cách tính. Khi đã có mẫu cụ thể $(x_1,...,x_n)$, kỳ vọng mẫu cũng nhận giá trị cụ thể, đó là

$$\overline{x} = \frac{1}{n}(x_1 + ... + x_n).$$

Nếu mẫu cụ thể cho dưới dạng bảng phân bố, chẳng hạn giá trị x_1 nhận n_1 lần, …, x_k nhận n_k lần, $n = \sum_{i=1}^{k} n_i$ thì

$$\overline{x} = \frac{(x_1 + ... + x_1) + ... + (x_k + ... + x_k)}{n} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i. \qquad (4.1.5)$$

Nếu mẫu cụ thể cho dưới dạng bảng phân bố ghép lớp, chỉ việc lấy giá trị trung tâm của mỗi lớp làm đại diện cho lớp đó

Ví dụ 4.3. Tuổi thọ (đơn vị: 100 giờ) một loại linh kiện do công ty A sản xuất ra được kiểm tra ngẫu nhiên, kết quả ghi thành bảng.

Tuổi	≤ 7	$7,0 \div 7,5$	$7,5 \div 8,0$	$8 \div 8,5$	$8,5 \div 9,0$	$\ge 9,0$
n_i	3	15	18	14	20	4

Kích thước mẫu là: $n = 3 + 15 + 18 + 14 + 20 + 4 = 74$.

Dùng giá trị chính giữa của mỗi khoảng làm đại diện cho khoảng đó, riêng với khoảng đầu và cuối ta chọn giá trị hợp lý nào đó, ví dụ 6,5 và 9,5 tương ứng, ta được

$$\overline{x} = \frac{1}{74}\left[3.6,5 + 15.7,25 + 18.7,75 + 14.8,25 + 20.8,75 + 4.9,5\right] = 8,057.\#$$

c) Phương sai mẫu

Định nghĩa. Ta gọi đại lượng

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i)^2 - (\overline{X})^2 \qquad (4.1.6)$$

là phương sai mẫu ứng với mẫu $(X_1,...,X_n)$ từ BNN gốc X.

Tính chất. Giả sử $E[X] = \mu; V[X] = \sigma^2$. Khi đó

$$E[S^2] = \frac{n-1}{n}\sigma^2. \qquad (4.1.7)$$

Từ (4.1.7) ta thấy $E[S^2] \ne \sigma^2$. Theo ngôn ngữ của lý thuyết ước lượng (ƯL) nêu ra ở bài §4.2 dưới đây, S^2 là một ƯL chệch của σ^2. Để có ƯL không chệch, ta phải sửa đổi S^2 đi đôi chút.

Định nghĩa. Phương sai mẫu hiệu chỉnh ứng với mẫu $(X_1,...,X_n)$, ký hiệu bởi $\tilde{S}^2$ là đại lượng xác định theo công thức

$$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \qquad (4.1.8)$$

Rõ ràng là $\tilde{S}^2 = \dfrac{n}{n-1}S^2.$

Tính chất

$$E[\tilde{S}^2] = \sigma^2. \qquad (4.1.9)$$

Định nghĩa. Căn bậc hai của phương sai mẫu (tương ứng, phương sai hiệu chỉnh mẫu) được gọi là độ lệch tiêu chuẩn mẫu (tương ứng, độ lệch tiêu chuẩn mẫu hiệu chỉnh).

Cách tính. Nếu mẫu nhận giá trị cụ thể $(x_1,...,x_n)$ thì phương sai mẫu và phương sai hiệu chỉnh mẫu sẽ nhận giá trị cụ thể s^2 và $(\tilde{s})^2$ với

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

$$(\tilde{s})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \qquad (4.1.10)$$

Tuy nhiên, thường thì mẫu cho dưới dạng bảng phân bố thực nghiệm, ở đó giá trị x_i nhận n_i lần, $n = n_1 + ... + n_k$. Khi đó

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

$$(\tilde{s})^2 = \frac{1}{n-1}\sum_{i=1}^{k}n_i(x_i - \overline{x})^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{k}n_i x_i^2 - (\overline{x})^2\right]. (4.1.11)$$

- ***Sử dụng máy tính bỏ túi***

Mọi phần mềm thống kê đều có chương trình tính kỳ vọng mẫu $\overline{x}$ và các phương sai $s^2, \tilde{s}^2$. Ở đây chúng ta đưa ra cách tính đơn giản dùng máy tính bỏ túi CASIO Fx-500MS.

Xoá nhớ thống kê `SHIFT` `MODE` `1` `=`

Gọi chương trình tính `MODE` `SD [2]`

Nhập dữ liệu. Chẳng hạn, ta cần nhập dữ liệu ở Ví dụ 4.3.

Giá trị 6,5 được tính ba lần `6.5` `M+` `M+` `M+`

Thay cho ấn M+ ba lần, ta có thể nhập theo cách thứ hai sau đây

`6.5` `SHIFT` `;` `3` `M+`

Cứ thế ta nhập cho hết dữ liệu.

Gọi kết quả. Nhập dữ liệu xong thì gọi kết quả như sau.

Giá trị có thể tính	Ấn	Kết quả cụ thể
$\sum x_i^2$	$\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{SUM}}$ $\boxed{\sum x^2\,[1]}$ $\boxed{=}$	4 841,43
$\sum x_i$	$\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{SUM}}$ $\boxed{\sum x\,[2]}$ $\boxed{=}$	596,25
$\overline{x}$	$\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{VAR}}$ $\boxed{\overline{x}[1]}$ $\boxed{=}$	8,0574
s	$\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{VAR}}$ $\boxed{x\sigma n\ [2]}$ $\boxed{=}$	0,7090
$\tilde{s}$	$\boxed{\text{SHIFT}}$ $\boxed{\text{S}-\text{VAR}}$ $\boxed{x\sigma n-1\ [3]}$ $\boxed{=}$	0,7138

§4.2. ƯỚC LƯỢNG ĐIỂM (1 tiết)

4.2.1. Những vấn đề chung về ước lượng tham số

a) Khái niệm ước lượng điểm

Định nghĩa. * Mỗi hàm (đo được) $T(X_1,...,X_n)$ của mẫu $(X_1,...,X_n)$ được gọi là một thống kê.

* Một thống kê $\hat{\theta} = T(X_1,...,X_n)$ dùng để ƯL (xấp xỉ) tham số θ được gọi là ƯL điểm của θ.

Sau này để cho gọn, ta sẽ gọi ƯL điểm là ƯL (estimator).

Nhận xét. Tham số không thể có mặt trong biểu thức của $T(X_1,...,X_n)$.

Khi mẫu đã nhận giá trị cụ thể $(x_1,...,x_n)$ thì ƯL cũng nhận giá trị cụ thể $T(x_1,...,x_n)$ (còn gọi là giá trị quan sát của ƯL) (estimate).

b) Các tiêu chuẩn ước lượng.

b1. Ước lượng không chệch

Định nghĩa. ƯL $\hat{\theta} = T(X_1,...,X_n)$ của tham số θ được gọi là không chệch nếu

$$E_\theta[T(X_1,...,X_n)] = \theta, \quad \theta \in \Theta. \qquad (4.2.1)$$

Trái lại nếu tồn tại $\theta \in \Theta$ để $E[\hat{\theta}] \neq \theta$ thì ƯL gọi là chệch. Độ chệch cho bởi

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta = E[\hat{\theta} - \theta]. \qquad (4.2.2)$$

Ước lượng không chệch đảm bảo rằng, nếu dùng ƯL này ta sẽ không mắc phải sai số hệ thống.

b2. Ước lượng vững. Một trong những đặc tính ưa chuộng của ƯL là khi kích thước mẫu n đủ lớn, ƯL sẽ có độ tin cậy đủ tốt.

Định nghĩa. Ước lượng $\hat{\theta}_n = T(X_1,...,X_n)$ được gọi là vững (consistant estimator) nếu với mọi $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1. \qquad (4.2.3)$$

Nói cách khác, khi n dần ra vô hạn, $\hat{\theta}_n$ dần đến θ theo xác suất.

Nói ngắn gọn, có thể yên tâm dùng xấp xỉ $\theta \approx \hat{\theta}_n$ khi n đủ lớn.

b3) Ước lượng hiệu quả

Định nghĩa. ƯL không chệch $\hat{\theta} = T(X_1,...,X_n)$ sao cho có phương sai nhỏ nhất trong các ƯL không chệch được gọi là ƯL không chệch với phương sai bé nhất (hay ƯL hiệu quả).

4.2.2. Ước lượng điểm các tham số thông dụng

a) Ước lượng giá trị trung bình

Định lý 4.2. Nếu BNN gốc X có $E[X] = \mu$, $V[X] = \sigma^2 < \infty$ thì kỳ vọng mẫu $\overline{X}$ là ƯL không chệch, vững và hiệu quả của μ.

Bảng 4.1. Một số ƯL tham số thông dụng

Tham số cần ƯL	**Thống kê ƯL**	$E_\theta[\hat{\theta}]$	$V_\theta[\hat{\theta}]$	**Tính chất của ƯL**
Kỳ vọng $\mu = E[X]$	$\overline{X} = \frac{1}{n}(X_1 + ... + X_n)$	μ	$\frac{\sigma^2}{n}$	Không chệch, vững, hiệu quả, hợp lý cực đại
Xác suất (tỷ lệ) p	$f = \frac{m}{n}$	p	$\frac{p(1-p)}{n}$	Không chệch, vững, hiệu quả, hợp lý cực đại
Phương sai $\sigma^2 = V[X]$	$\tilde{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})$	σ^2	…	Không chệch, vững
	$S^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$	σ^2	…	Không chệch, vững
	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$	$\frac{n-1}{n}\sigma^2$	…	Chệch, vững, hợp lý cực đại

4.2.3. Phương pháp hợp lý cực đại (☼)

b) Thảo luận	
c) Tự học	4.2.3. Phương pháp hợp lý cực đại
d) Bài tập chuẩn bị tối thiểu	**Bài tập về nhà cho cả chương 4** Tài liệu [1]: 4(1 – 4 – 5 – 6 – 10 – 11 – 12 – 13 – 14 – 17 – 19 – 21 – 23 – 24 – 25(a) – 26(a,b) – 27 – 29 – 30 – 31 – 32 –33- 34 – 35 – 37). Tài liệu [2]: Tr 153-157: 11, 12,15,17, 19,22
Tài liệu	Tài liệu [1], tr ....

Bài giảng 10: Kiểm định giả thuyết và ước lượng tham số (tiếp)

Chương, mục: 4

Tiết thứ: 37-40 Tuần thứ: 10

Mục đích, yêu cầu:

- Cách dùng, cách tìm khoảng tin cậy
- Ba bài toán kiểm định thông dụng

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

Kiểm tra phần xác suất

§4.3. Khoảng tin cậy

§4.4. Bài toán kiểm định giả thuyết tổng quát

KIỂM TRA phần xác suất (1 tiết)

§4.3. KHOẢNG TIN CẬY (1tiết)

4.3.1. Khái niệm về khoảng tin cậy

Định nghĩa. Giả sử $(X_1,...,X_n)$ là một mẫu ngẫu nhiên từ BNN gốc X mà luật phân bố phụ thuộc vào tham số thực $\theta \in \Theta$. Khoảng $(\theta_1; \theta_2)$, trong đó $\theta_1 = T_1(X_1,...,X_n); \theta_2 = T_2(X_1,...,X_n)$ là hai thống kê nào đó được gọi là khoảng tin cậy với độ tin cậy $100\beta\%$ của θ nếu

$$P(\theta_1 < \theta < \theta_2) = \beta. \qquad (4.3.1)$$

Đặt $\alpha = 1-\beta$, gọi là mức ý nghĩa.

Hiệu $\theta_2 - \theta_1$ được gọi là độ rộng khoảng tin cậy.

Khoảng tin cậy được gọi là đối xứng nếu

$$P(\theta \le \theta_1) = P(\theta \ge \theta_2).$$

Thường thì người ta chọn $\beta = 0,9; 0,95; 0,98; 0.99...$

Khi có mẫu cụ thể $(x_1, ... , x_n)$ thay vào trên ta được khoảng tin cậy cụ thể (a; b) = $(\theta_1(x_1, ... , x_n); \theta_2(x_1, ... , x_n))$ nào đó. Kể cả khi đó ta cũng không được viết $P(a < \mu < b) = \beta$!

4.3.2. Khoảng tin cậy cho kỳ vọng μ của phân bố chuẩn $N(\mu;\sigma^2)$

a) Đã biết phương sai σ^2. Lưu ý rằng khi σ^2 là giá trị đã biết, nó không còn là tham số. Gọi $\overline{X}$ là trung bình mẫu của mẫu $(X_1,...,X_n)$,

Ta biết rằng $Z = (\overline{X} - \mu)\sqrt{n} / \sigma \sim N(0;1)$. Suy ra

$$P\left(\frac{\left|\overline{X}-\mu\right|}{\sigma}\sqrt{n} < z_{\frac{\alpha}{2}}\right) = P\left(|Z| < z_{\alpha/2}\right) = 1-2P\left(|Z| \geq z_{\alpha/2}\right) = 1-2\,\frac{\alpha}{2} = \beta$$

$$\Leftrightarrow P\left(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2} < \mu < \overline{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right) = \beta. \quad (4.3.2)$$

Theo định nghĩa, khoảng tin cậy đối xứng với độ tin cậy $\beta = 1-\alpha$ của μ là

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2};\ \overline{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right). \qquad (4.3.3)$$

Người ta hay viết khoảng này dưới dạng $\overline{X} \pm \frac{\sigma}{\sqrt{n}}z_{\alpha/2}$ hoặc

$$\overline{X} \pm \varepsilon \quad \text{với} \quad \varepsilon = \frac{\sigma}{\sqrt{n}}z_{\alpha/2}. \qquad (4.3.3')$$

Kích thước mẫu

Độ rộng khoảng tin cậy $I = 2\varepsilon$, độ tin cậy β và kích thước mẫu n liên hệ chặt chẽ với nhau qua phương trình (4.3.3'). Nếu biết 2 trong 3 đại lượng đó ta có thể tính được đại lượng còn lại.

Bây giờ giả sử với độ tin cậy β đã cho, hãy tìm kích thước mẫu n tối thiểu để độ rộng khoảng tin cậy không vượt quá giá trị I_0 cho trước. Muốn vậy

$$I_0 = 2\varepsilon_0 \geq I = 2\varepsilon = 2.\frac{\sigma}{\sqrt{n}}z_{\alpha/2} \Leftrightarrow n \geq \frac{4\sigma^2 z_{\alpha/2}^2}{I_0^2}.$$

Từ đó ta tính được kích thước mẫu n cần thiết là

$$n \geq n_0 = \frac{4\sigma^2}{I_0^2}z_{\alpha/2}^2 = \frac{\sigma^2}{\varepsilon_0^2}z_{\alpha/2}^2. \qquad (4.3.4)$$

trong đó I_0 (ε_0) là độ rộng (nửa độ rộng) khoảng tin cậy đòi hỏi.

Ví dụ 4.8. Xem khối lượng một loại sản phẩm là BNN chuẩn với độ lệch chuẩn là 1 gam. Chọn ngẫu nhiên 14 sản phẩm, đem cân thử ta được kết quả

Trọng lượng	18	19	20	21
Số sản phẩm	1	5	6	2

Tìm khoảng tin cậy đối xứng 95% của khối lượng trung bình của loại sản phẩm trên. Nếu muốn độ rộng khoảng tin cậy 95% là 0,4g, ta phải cân thử thêm bao nhiêu sản phẩm nữa?

Khối lượng $X \sim N(\mu;1)$; $1-\alpha = 0,95$. $\alpha/2 = 0,025$;

$$z_{\alpha/2} = z_{0,025} = 1,960;\ \overline{x} = \frac{1}{14}(1.18+5.19+6.20+2.21) = 19,643$$

$$\Rightarrow \varepsilon = \frac{1}{\sqrt{14}}1,960 = 0,524.$$

Như vậy, qua mẫu cụ thể ta tính được khoảng tin cậy đối xứng của khối lượng trung bình là $19,643 \pm 0,524 = (19,12;\ 20,17)$.

Bây giờ ta tìm số lần cân tối thiểu. Ta có

$$n \approx \frac{4\sigma^2}{I_0^2}z_{\alpha/2}^2 = \frac{4.1^2}{0,4^2}1,960^2 = 96.04. \text{ Chọn } n_0 = 96. \qquad \#$$

Thực tế, khi tăng n từ $2 \to 3 \to 4$, độ rộng khoảng tin cậy giảm mạnh (tức là độ chính xác tăng mạnh). Nếu tiếp tục tăng n thì nửa độ dài khoảng tin cậy tính theo (4.3.3') có thể sẽ nhỏ hơn độ chệch δ và độ chính xác có thể không giảm nhiều so với nỗ lực bỏ ra (xem (4.2.4)).

Từ đó trong kỹ thuật, thường phải tiến hành **ít ra là 2** thí nghiệm; đồng thời, chỉ những thí nghiệm đặc biệt người ta mới tiến hành **quá 4**. Nếu muốn cải thiện sai số thì nên nghĩ đến thay dụng cụ hoặc phương pháp đo hoặc cả hai làm cho sai số hệ thống (độ chệch) nhỏ hơn.

Tóm tắt ƯL khoảng cho ở Bảng 4.2.

Bảng 4.2. Tóm tắt ƯL khoảng

Tham số cần ƯL	**Từ phân bố**	**Thông tin bổ sung**	**Khoảng tin cậy đối xứng**
Kỳ vọng μ	Chuẩn $N(\mu;\sigma^2)$	σ^2 đã biết	$\left(\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$
		σ^2 chưa biết	$\left(\overline{X} \pm \frac{\tilde{S}}{\sqrt{n}} t_{\alpha/2}(n-1)\right)$
	Bất kỳ	Mẫu lớn	$\left(\overline{X} \pm \frac{S}{\sqrt{n}} z_{\alpha/2}\right)$
	Mũ $E(\lambda)$	Mẫu lớn	$\left(\frac{\overline{X}}{1+z_{\alpha/2}/\sqrt{n}}; \frac{\overline{X}}{1-z_{\alpha/2}/\sqrt{n}}\right)$
	Poisson $P(\lambda)$	Mẫu lớn	$(\lambda_1;\lambda_2)$: $\lambda_1;\lambda_2$ thoả mãn $(\lambda-\overline{X})^2 = z_{\alpha/2}^2\lambda/n$
Xác suất p	Nhị thức $B(1;p)$	Mẫu lớn	$f \pm z_{\alpha/2}\sqrt{\frac{f(1-f)}{n}}$
Phương sai σ^2	Chuẩn $N(\mu;\sigma^2)$	μ đã biết	$\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{\alpha/2}^2(n)}; \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi_{(1-\alpha/2)}^2(n)}\right)$
		μ chưa biết	$\left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi_{\alpha/2}^2(n-1)}; \frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\chi_{(1-\alpha/2)}^2(n-1)}\right)$

Ví dụ 4.9. Để xác định khối lượng trung bình của các bao gạo trong kho người ta cân ngẫu nhiên 15 bao và thu được $\overline{x} = 49,8$ và độ lệch chuẩn là 0,397.

Hãy xác định khoảng tin cậy (đối xứng) của khối lượng trung bình với độ tin cậy 95% khi coi khối lượng của các bao gạo có phân bố chuẩn trong các trương hợp:

i) Biết phương sai là 0,5;

ii) Chưa biết phương sai.

Giải. i) $\alpha/2 = (1-0,95)/2 = 0,025$; $U_{0,025} = 1,960$;

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{\alpha/2} = \frac{\sqrt{0,5}}{\sqrt{15}} 1,960 = 0,358 \Rightarrow \overline{X} \pm \varepsilon = (49,44;\ 50,16).$$

ii) Theo (4.3.6), $t_{0,025}(15-1) = 2,145$, $\varepsilon = \frac{0,397}{\sqrt{15}} 2,145 = 0,220$.

Khoảng tin cậy 95% qua mẫu cụ thể là $49,80 \pm 0,22 = (49,58;\ 50,02)$. #

Ví dụ 4.10. Khi xác định Niken trong thép bằng phương pháp khối lượng người ta tìm được các giá trị 4,64; 4,67; và 4,65 (%) Ni.

i) Xác định khoảng tin cậy 95% của hàm lượng Niken trung bình.

ii) Thực ra, phương pháp khối lượng để xác định hàm lượng Ni trong thép là phương pháp khá tốt: Theo các cuốn sổ tay luyện kim, sai số tương đối của phương pháp này với hàm lượng Ni từ 2% đến 10% là 0,5%; ở đây, $\sigma = \frac{0,5}{100} 4,65(\%) = 0,023(\%)$. Tìm khoảng tin cậy 95% cho hàm lượng Ni trung bình.

$(4.653 \pm 0,038)$; $(4,653 \pm 0,026)$,

Ví dụ 4.13. Một vùng có 2000 hộ gia đình. Để điều tra nhu cầu tiêu dùng một loại hàng nào đó, người ta chọn ngẫu nhiên 100 hộ và thấy có 60 hộ có nhu cầu về loại hàng trên. Với độ tin cậy 95%, hãy ƯL bằng khoảng tin cậy (đối xứng) số hộ gia đình có nhu cầu về loại hàng trên.

Giải. Gọi M là số gia đình có nhu cầu. Tỷ lệ số hộ có nhu cầu là $p = M/2000$.

$p \in (0,6 \pm 0,096) = (0,504;\ 0,696)$

$\Rightarrow M \in (1008;\ 1393)$.

Nhận xét. i) *Trường hợp n nhỏ.*

ii) *Trường hợp p nhỏ.* (Xem [1]).

Ví dụ 4.14. Nguồn điện thế V được đo 6 lần và các kết quả đo là

221,5; 220,6; 219,4; 218,5; 221,3; 221,0 (V).

Giả sử phép đo không có sai số hệ thống và được mô hình hoá bởi BNN $X = V + \nu$ với giả thiết rằng ν có phân bố chuẩn $N(0;\sigma^2)$. Tìm khoảng tin cậy 95% cho σ^2 trong các trường hợp:

a) $V = 220\,(V)$ đã biết, b) V chưa biết.

Giải. $\left(0,547;\ 6,395\right)$ và cho σ là (0,74; 2,53). (Khoảng này lệch so với ƯL $s^{*2} = 1,318$ của σ^2).

b) ƯL khoảng 95% cho σ^2: $(0,548;\ 8,457)$. #

§4.4. BÀI TOÁN KIỂM ĐỊNH GIẢ THUYẾT TỔNG QUÁT (1 tiết)

4.4.1. Giả thuyết, đối thuyết, quy tắc quyết định, miền bác bỏ

Giả sử ta biết chắc chắn rằng BNN X có phân bố thuộc vào một họ ₤ nhưng chúng ta lại không biết, và muốn kiểm tra xem phải chăng phân bố của X thuộc vào họ con $₤_H$ hay họ con $₤_K = ₤ - ₤_H$.

Ta gọi mệnh đề

"Phân bố của X thuộc vào họ $₤_H$" là giả thuyết, ký hiệu là H_0;

"Phân bố của X thuộc vào họ $₤_K$" là đối thuyết, ký hiệu là H_1.

Để trả lời câu hỏi đặt ra, ta lập một mẫu quan sát $(X_1,...,X_n)$ về X. Tiếp theo ta xét BNN $T = T(X_1,...,X_n)$ (với $T(x_1,...,x_n)$ là hàm đo được) và gọi $\mathcal{H} \subset \mathbb{R}$ là tập giá trị của nó. Chia $\mathcal{H}$ làm hai phần R và $\overline{R} = \mathcal{H} - R$ (R đo được) rồi quyết định theo quy tắc sau đây:

Nếu $T(X_1,...,X_n) \in R$ thì bác bỏ giả thuyết H_0, chấp nhận đối thuyết H_1.

Nếu $T(X_1,...,X_n) \in \overline{R}$ thì chấp nhận giả thuyết H_0 (đúng ra, ta chưa có cơ sở để bác bỏ H_0, cần chấp nhận H_0 cho đến khi có thông tin mới).

BNN $T = T(X_1,...,X_n)$ được gọi là thống kê kiểm định, tập R được gọi là miền bác bỏ (hoặc miền tiêu chuẩn, miền tới hạn) (critical region); còn tập $\overline{R}$ được gọi là miền chấp nhận (region of acceptance) của giả thuyết H_0.

Việc lựa chọn quy tắc quyết định tương đương với việc lựa chọn thống kê kiểm định $T = T(X_1,...,X_n)$ và miền bác bỏ R. Quy trình dẫn đến quyết định về giả thuyết được gọi là kiểm định giả thuyết (test of a hypothesis), gọi tắt là kiểm định (test).

4.4.2. Hai loại sai lầm

Mỗi kiểm định đều dẫn đến một trong hai loại sai lầm sau đây:

Sai lầm loại I: Thực ra H_0 là đúng nhưng ta lại bác bỏ;

Sai lầm loại II: Thực ra H_0 là sai nhưng ta lại chấp nhận.

Để cho rõ, ta đưa ra bảng sau nêu lên các tình huống gặp phải.

Kết luận / Thực tế	Chấp nhận H_0 (bác bỏ H_1)	Bác bỏ H_0 (chấp nhận H_1)
H_0 đúng	Kết luận đúng	Sai lầm (loại I)
H_0 sai	Sai lầm (loại II)	Kết luận đúng

Ký hiệu

$$P(R|H_0) = P\left(T(X_1,...,X_n) \in R|H_0\right) \qquad (4.4.1)$$

là xác suất (mắc) sai lầm loại I. Lưu ý rằng H_0 tham gia vào ký hiệu này không có nghĩa là xác suất điều kiện, đơn giản, nó nói lên rằng H_0 là đúng. Tương tự, xác suất (mắc) sai lầm loại II xác định bởi

$$P(\overline{R}|H_1) = P\left(T(X_1,...,X_n) \notin R|H_1\right) \qquad (4.4.2)$$

4.4.3. Lựa chọn kiểm định

Cái mà ta có thể làm sau khi đã lựa chọn thống kê kiểm định là lựa chọn miền bác bỏ R. Lẽ tự nhiên, ta mong muốn R làm cực tiểu cả 2 loại xác suất phạm sai lầm. Tuy nhiên điều này nói chung không thể thực hiện được. Thực vậy, vì

$$P(R|H_0) + P(\overline{R}|H_0) = 1;\ \ P(R|H_1) + P(\overline{R}|H_1) = 1;$$

miền bác bỏ R cực tiểu $P(R|H_0)$ thường sẽ làm tăng $P(\overline{R}|H_1)$.

Giả thuyết là điều chính yếu. So sánh mức độ nguy hại của sai lầm loại I với sai lầm loại II, người ta có xu hướng coi trọng xác suất mắc sai lầm loại I, nó không được lớn. Từ đó, cho trước $\alpha \in (0;1)$, người ta chỉ xét những kiểm định mức α, với nó $P(R|H_0) \le \alpha$. Bài toán đặt ra là, chọn R sao cho

$$\begin{cases} P(R|H_0) \le \alpha; \\ \beta = P(\overline{R}|H_1) \ \text{đạt cực tiểu.} \end{cases} \qquad (4.4.3)$$

Bởi vì $P(R|H_1) = 1 - P(\overline{R}|H_1)$ nên $P(\overline{R}|H_1)$ đạt cực tiểu khi và chỉ khi $P(R|H_1)$ đạt cực đại; bài toán (4.4.3) tương đương với

$$\begin{cases} P(R|H_0) \le \alpha; \\ P(R|H_1) \ \text{đạt cực đại.} \end{cases} \qquad (4.4.3')$$

Giá trị α gọi là mức ý nghĩa, thường được chọn là 0,10; 0,05; 0,02; 0,01 … tuỳ vào bài toán cụ thể (*).

4.4.4. Các bài toán kiểm định tham số cơ bản

Thông thường, họ phân bố £ được tham số hoá, nó được cho bởi họ hàm phân bố đã biết dạng $F(x, \theta)$, nhưng tham số $\theta \in \Theta \subset \mathbb{R}^k$ lại chưa biết. Giả thuyết H_0 bây giờ trở thành $\theta \in \Theta_{H_0} \subset \Theta$; đối thuyết H_1 trở thành $\theta \in \Theta_{H_1} = \Theta - \Theta_{H_0}$.

Nếu Θ_H gồm đúng một điểm $\{\theta_0\}$ thì ta có giả thuyết đơn; trái lại, ta có giả thuyết hợp. Tương tự như vậy ta có khái niệm đối thuyết đơn và đối thuyết hợp.

Nhiều bài toán kiểm định giả thuyết hợp có thể chuyển về giả thuyết đơn. Để đơn giản ta chỉ trình bày trường hợp giả thuyết đơn.

Định nghĩa. Đối với bài toán kiểm định giả thuyết đơn $H_0 : \theta = \theta_0$, xét các quy tắc kiểm định mức α: $P(R|H_0) \le \alpha$. Ký hiệu xác suất để thống kê kiểm định $T(X_1, ..., X_n)$ rơi vào miền bác bỏ R tính với phân bố $F(x, \theta)$ là $P(R|\theta)$. Đặt

$$\beta(\theta) = P(R|\theta), \quad \theta \in \Theta_{H_1} \qquad (4.4.4)$$

gọi là hàm lực lượng (hàm mạnh) của kiểm định.

Nếu $\theta_1 \in \Theta_{H_1}$ cố định mà $\beta(\theta_1)$ đạt cực đại thì kiểm định được gọi là mạnh nhất mức α.

Nếu kiểm định mức α nói trên trên không phụ thuộc vào $\theta_1 \in \Theta_{H_1}$ thì nó được gọi là kiểm định mạnh đều nhất mức α.

Chúng ta còn hạn chế bài toán hơn nữa, chỉ xét trường hợp có một tham số. Khi ấy, ba dạng sau đây được quan tâm đặc biệt.

Bài toán I - Kiểm định hai phía.

$$H_0 : \theta = \theta_0 \ / \ H_1 : \theta \ne \theta_0$$

Bài toán II - Kiểm định một phía (phía phải).

$$H_0 : \theta = \theta_0 \; / \; H_1 : \theta > \theta_0$$

Bài toán III - Kiểm định một phía (phía trái).

$$H_0 : \theta = \theta_0 \; / \; H_1 : \theta < \theta_0 .$$

Các kiểm định nói đến sau đây đều là mạnh nhất, mạnh đều nhất, tối ưu hay xấp xỉ như vậy và đã được chứng minh chặt chẽ về mặt toán học. Tuy nhiên các chứng minh khá phức tạp, độc giả có thể tham khảo ở [2], [9], [11], chúng ta chỉ dẫn ra và xử lý sơ bộ để minh hoạ ý tưởng chính.

BÀI TẬP: Ước lượng điểm (1 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 11: Kiểm định giả thuyết và ước lượng tham số (tiếp)

Chương, mục: 4

Tiết thứ: 41-44 Tuần thứ: 11

Mục đích, yêu cầu:

- Nắm được bảng tóm tắt để làm được một số bài tập thực tế

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§4.5. Các kiểm định dùng một mẫu

Bài tập về ước lượng khoảng

§4.5. CÁC KIỂM ĐỊNH DÙNG MỘT MẪU (2 tiết)

4.5.1. Kiểm định giả thuyết về giá trị trung bình

Giả sử có mẫu $(X_1,...,X_n)$ từ phân bố chuẩn $N(\mu;\sigma^2)$.

a) Phương sai σ^2 đã biết. Các bài toán nêu trên trở thành:

Bài toán I . $H_0: \mu=\mu_0 \ / \ H_1: \mu \neq \mu_0$

Bài toán II. $H_0: \mu=\mu_0 \ / \ H_1: \mu > \mu_0$

Bài toán III . $H_0: \mu=\mu_0 \ / \ H_1: \mu < \mu_0$.

Thống kê kiểm định được chọn là

$$Z=\frac{\overline{X}-\mu_0}{\sigma}\sqrt{n}.$$

Đối với Bài toán I, nếu H_0 là đúng thì $\overline{X}$ sẽ không quá xa μ_0. Vì thế, ta nên chọn kiểm định sao cho bác bỏ H_0 nếu $|Z|>c$. Để xác định hằng số c ta thấy nếu H_0 đúng thì $Z \sim N(0;1)$; quy tắc có mức α, vậy chọn c từ điều kiện

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

Như vậy, với bài toán I, ta chọn quyết định bác bỏ H_0 nếu

$$|Z|=\frac{|\overline{X}-\mu_0|}{\sigma}\sqrt{n}>z_{\alpha/2} \qquad (4.5.1)$$

trong đó z_α là phân vị chuẩn mức α.

Thực ra có thể chứng minh chặt chẽ rằng, đối với Bài toán I, kiểm định vừa xây dựng là mạnh đều nhất mức α (xem [2], tr. 192).

Tương tự, với hai bài toán còn lại, ta sẽ bác bỏ H_0 khi và chỉ khi

Với Bài toán II: $$Z=\frac{\overline{X}-\mu_0}{\sigma}\sqrt{n}>z_\alpha; \qquad (4.5.2)$$

Với Bài toán III: $\quad Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} < z_{1-\alpha} = -z_{\alpha}$. (4.5.3)

Bảng 4.3. Tóm tắt kiểm định: Mẫu $(X_1,...,X_n)$, *mức ý nghĩa* α

<table>
<tr><th colspan="2">Tham số cần KĐ</th><th rowspan="2">Từ phân bố</th><th colspan="2">Bác bỏ H_0 khi</th></tr>
<tr><th>Giả thuyết</th><th>Đối thuyết</th><th>σ^2 đã biết</th><th>σ^2 chưa biết</th></tr>
<tr><td rowspan="3">Kỳ vọng $\mu = \mu_0$</td><td>$\mu \neq \mu_0$</td><td rowspan="3">Chuẩn $N(\mu;\sigma^2)$</td><td>$|Z| = \dfrac{|\overline{X} - \mu_0|}{\sigma}\sqrt{n} > z_{\alpha/2}$</td><td>$|T| = \dfrac{|\overline{X} - \mu_0|}{\tilde{S}}\sqrt{n} > t_{\alpha/2}(n-1)$</td></tr>
<tr><td>$\mu > \mu_0$</td><td>$Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} > z_{\alpha}$</td><td>$T = \dfrac{\overline{X} - \mu_0}{\tilde{S}}\sqrt{n} > t_{\alpha}(n-1)$</td></tr>
<tr><td>$\mu < \mu_0$</td><td>$Z = \dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} < -z_{\alpha}$</td><td>$T = \dfrac{\overline{X} - \mu_0}{\tilde{S}}\sqrt{n} < -t_{\alpha}(n-1)$</td></tr>
<tr><td rowspan="3">Xác suất $p = p_0$</td><td>$p \neq p_0$</td><td rowspan="3">Không-một $B(1;p)$</td><td colspan="2">$|Z| = \dfrac{|f - p_0|}{\sqrt{p_0(1-p_0)}}\sqrt{n} > z_{\alpha/2}$</td></tr>
<tr><td>$p > p_0$</td><td colspan="2">$Z = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} > z_{\alpha}$</td></tr>
<tr><td>$p < p_0$</td><td colspan="2">$Z = \dfrac{f - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} < z_{1-\alpha} = -z_{\alpha}$</td></tr>
<tr><td rowspan="3">Phương sai $\sigma^2 = \sigma_0^2$</td><td>$\sigma^2 \neq \sigma_0^2$</td><td rowspan="3">$N(\mu;\sigma^2)$</td><td colspan="2">$\left[\begin{array}{l} \chi^2 = \dfrac{(n-1)\tilde{S}^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}(n-1) \\ \chi^2 = \dfrac{(n-1)\tilde{S}^2}{\sigma_0^2} > \chi^2_{\alpha/2}(n-1) \end{array}\right.$</td></tr>
<tr><td>$\sigma^2 > \sigma_0^2$</td><td colspan="2">$\chi^2 = \dfrac{(n-1)\tilde{S}^2}{\sigma_0^2} > \chi^2_{\alpha}(n-1)$</td></tr>
<tr><td>$\sigma^2 < \sigma_0^2$</td><td colspan="2">$\chi^2 = \dfrac{(n-1)\tilde{S}^2}{\sigma_0^2} < \chi^2_{1-\alpha}(n-1)$</td></tr>
</table>

Chú ý. Khi n lớn ($n > 30$), không cần phân bố chuẩn và $t_{\alpha}(n) \approx z_{\alpha}$.

Ví dụ 4.15. Ta lấy lại Ví dụ 4.14 ở đó có tập gồm 6 số liệu điện thế với $\overline{x} = 220,38;\ \tilde{s}^2 = 1,4056$. Ta muốn kiểm định giả thuyết với mức ý nghĩa 0,05 rằng điện thế trung bình là 220V khi

a) Biết phương sai của các sai số đo là 1;

b) Chưa biết phương sai.

Giải. $H_0 : \mu = 220\ /\ H_1 : \mu \neq 220;\ \alpha = 0,05$.

a) $\quad |Z| = \dfrac{|220,38 - 220|}{1}\sqrt{6} = 0,93$.

$z_{\alpha/2} = z_{0,025} = 1,960$; $|Z| < z_{\alpha/2}$: Không bác bỏ H_0, coi điện thế trung bình không khác 220V một cách có ý nghĩa.

b) $\quad |T| = \dfrac{|220,38 - 220|}{1,1856}\sqrt{6} = 0,785$.

$t_{\alpha/2}(n-1) = t_{0,025}(5) = 2.571$; $|T| < t_{\alpha/2}(n-1)$: Ta không bác bỏ H_0, coi điện thế trung bình bằng 220V. #

Bảng 4.3. Tóm tắt kiểm định: Mẫu $(X_1,...,X_n)$*, mức ý nghĩa* α

Ví dụ 4.16. Quá trình sản xuất một loại vòng bi được coi là bình thường nếu khối lượng của vòng bi có phân bố chuẩn với trung bình 5 và độ lệch chuẩn 0,1 (ounce). Sau khi có cải tiến sản xuất, giám đốc kỹ thuật nghi ngờ rằng khối lượng trung bình của vòng bi đã tăng lên (phương sai không đổi) mặc dầu vẫn có phân bố chuẩn. Một mẫu ngẫu nhiên 15 chiếc được lấy ra kiểm tra thì thấy khối lượng trung bình là 5,04. Đánh giá xem khối lượng trung bình đã tăng lên đáng kể hay chưa với mức ý nghĩa 0,05 và 0,10.

Giải. Xét bài toán $H_0: \mu = 5$ / $H_1: \mu > 5$. Ta có

$$Z = \frac{\overline{X} - \mu_0}{\sigma}\sqrt{n} = \frac{5,04 - 5,00}{0,1}\sqrt{15} = 1,549.$$

Tra bảng, $z_\alpha = z_{0,05} = 1,645$. Giá trị của thống kê Z chưa vượt ngưỡng ($Z < z_{0,05}$) nên ta không bác bỏ H_0: Coi khối lượng chưa tăng lên đáng kể.

Với $\alpha = 0,10$ thì $z_\alpha = z_{0,10} = 1,280$. Giá trị của thống kê Z đã vượt ngưỡng ($Z > z_{0,10}$) nên bác bỏ H_0: Dường như đã có trục trặc trong quá trình sản xuất, coi khối lượng trung bình đã tăng lên. #

4.5.4. Phương pháp P-giá trị

Các Ví dụ đó cho ta thấy nhược điểm của phương pháp quyết định đã nêu, gọi là phương pháp cổ điển. Các phương pháp quyết định thống kê ngày nay đều dùng P-giá trị (P-value).

Định nghĩa. Mức ý nghĩa nhỏ nhất tại đó giả thuyết H_0 bị bác bỏ được gọi là P-giá trị kết hợp với mẫu quan sát được.

Tính P-giá trị. Theo Ví dụ 4.16 ta đã tính ra được $\dfrac{\overline{X} - \mu_0}{\sigma}\sqrt{n} = 1,549$. Đối thuyết là $H_1: \mu > \mu_0 = 5$. Như vậy, giả thuyết bị bác bỏ ở mức α khi và chỉ khi $z_\alpha < 1,549 = z_{0,0607}$. Suy ra P-giá trị (một phía) (one-sised P-value) là 0,0607.

Nếu ta xét kiểm định hai phía (đối thuyết $H_1: \mu \neq \mu_0 = 5$), giả thuyết bị bác bỏ ở mức α khi và chỉ khi $z_{\alpha/2} < 1,549 = z_{0,1214/2}$. Vậy P-giá trị (hai phía) (two-sised P-value) là 0,1214.

Tổng quát, giả sử ta dùng kiểm định Z, tức là thống kê kiểm định Z là BNN có phân bố chuẩn tắc $N(0;1)$, và giả sử giá trị cụ thể của Z là z. Khi đó P-giá trị hai phía sẽ là giá trị P mà $|z| = z_{P/2}$ hay

$$P = 2(1 - \Phi(|z|)) = 1 - 2\Phi_0(|z|) \qquad (4.5.13)$$

trong đó $\Phi_0(x)$ là hàm Laplace.

Đối với kiểm định một phía, chẳng hạn phía phải $\mu > \mu_0$ và $z > 0$, P-giá trị một phía tương ứng sẽ là giá trị P mà $z = z_P$ hay

$$P = 1 - \Phi(z) = 1/2 - \Phi_0(z)\,. \qquad (4.5.14)$$

(Nếu z < 0, giả thuyết hiển nhiên bị bác bỏ, P-giá trị bằng 0).

Như vậy, P-giá trị hai phía gấp đôi P-giá trị một phía.

Đánh giá kết quả. Nhà thống kê tính ra P-giá trị, để độc giả tự đánh giá lấy kết quả. Hướng dẫn của nhà thống kê như sau. Giả sử mức ý nghĩa α đã chọn.

Nếu $P \ll \alpha$: Bác bỏ H_0 một cách mạnh mẽ;

$P \le \alpha$: Bác bỏ H_0;

$\alpha < P$: Chưa có cơ sở bác bỏ H_0;

$\alpha \ll P$: Chấp nhận H_0 khá chắc chắn.

Chẳng hạn, nếu không có gì đặc biệt thì hay chọn $\alpha = 0{,}05$ và

Nếu $0{,}05 \le P$: Không đủ cơ sở để bác bỏ H_0;

$0{,}01 < P < 0{,}05$: Đủ cơ sở để bác bỏ H_0;

$P < 0{,}01$: Có cơ sở mạnh, hùng hồn để bác bỏ H_0.

Ý tưởng chủ đạo là: P-giá trị càng nhỏ thì càng bác bỏ giả thuyết mạnh, P-giá trị càng lớn thì càng chấp nhận giả thuyết mạnh.

Tất cả các phần mềm thống kê ngày nay đều đưa ra P-giá trị cho mỗi bài toán kiểm định.

4.5.5. Các khía cạnh khác (☼)

a) Mối quan hệ giữa kiểm định giả thuyết và ước lượng khoảng

b) Xác định xác suất mắc sai lầm loại II

BÀI TẬP: Khoảng tin cậy (2 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 12: Kiểm định giả thuyết và ước lượng tham số (tiếp)

Chương, mục: 4

Tiết thứ: 45-48 Tuần thứ: 12

Mục đích, yêu cầu:

- Nắm được bảng tóm tắt để làm được một số bài toán thực tế
- Nắm được vài dạng áp dụng của tiêu chuẩn kiểm định Khi bình phương

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

§4.6. Các kiểm định dùng hai mẫu

§4.7. Kiểm định phi tham số

Bài tập về kiểm định giả thuyết

§ 4.6. CÁC KIỂM ĐỊNH DÙNG 2 MẪU (bài toán so sánh) (1 tiết)

Giả sử từ tổng thể thứ nhất với BNN gốc $X \sim N(\mu_1;\sigma_1^2)$ ta rút ra mẫu $(X_1,...,X_{n_1})$ và từ tổng thể thứ hai với BNN gốc $Y \sim N(\mu_2;\sigma_2^2)$ ta rút ra mẫu $(Y_1,...,Y_{n_2})$; hơn nữa hai mẫu này được rút ra một cách độc lập với nhau. Như trước đây, ta ký hiệu $\overline{X}, \overline{Y}, \tilde{S}_X^2, \tilde{S}_Y^2$ là kỳ vọng mẫu và phương sai mẫu hiệu chỉnh ứng với hai mẫu này.

4.6.1. So sánh hai giá trị trung bình

Xét bài toán kiểm định

$$H_0: \mu_1 - \mu_2 = D_0 \;/\; H_1: \mu_1 - \mu_2 \neq D_0$$

trong đó D_0 là một hằng số nào đó. Chúng ta sẽ trình bày trường hợp đặc biệt khi $D_0 = 0$, đó là

$$H_0: \mu_1 = \mu_2 \;/\; H_1: \mu_1 \neq \mu_2;$$

a) Trường hợp σ_1^2 và σ_2^2 đã biết.

Vì $\overline{X} \sim N\left(\mu_1;\dfrac{\sigma_1^2}{n_1}\right)$; $\overline{Y} \sim N\left(\mu_2;\dfrac{\sigma_2^2}{n_2}\right)$, chúng độc lập nên $\overline{X}-\overline{Y}$ có phân bố $N\left(\mu_1-\mu_2;\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}\right)$. Chuẩn hoá ta được

$$\left[\overline{X}-\overline{Y}-(\mu_1-\mu_2)\right] / \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \sim N(0;1). \qquad (4.6.1)$$

Nếu giả thuyết đúng thì $Z = (\overline{X} - \overline{Y}) / \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}} \sim N(0; 1)$.

Như trước đây, giả thuyết bị bác bỏ ở mức ý nghĩa α nếu

$$|Z| = \frac{|\overline{X} - \overline{Y}|}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} > z_{\alpha/2}. \qquad (4.6.2)$$

Ví dụ 4.20. Hai nhóm sản xuất phải xác định nitơ trong mẫu cho trước theo phương pháp vi phân tích (xinkhonin) đã tìm được các giá trị sau (% N):

Nhóm I: 9,29; 9,38; 9,35; 9,43 $\to \overline{x} = 9,363$; $\tilde{s}_X = 0,0582$

Nhóm II: 9,53; 9,48; 9,61; 9,68 $\to \overline{y} = 9,575$; $\tilde{s}_Y = 0,0881$.

Kiểm tra xem có sự khác biệt giữa 2 giá trị trung bình hay không.

Giải. Giả sử các mẫu được lấy từ phân bố chuẩn. Ta thấy $n_1 = n_2 = 4 < 30$. Dùng tiêu chuẩn F để so sánh hai phương sai thấy không có sự khác biệt lớn về mặt thống kê giữa σ_1^2 víi σ_2^2. Ta tiến hành kiểm định giả thuyết $\mu_1 = \mu_2$ với đối thuyết $\mu_1 \neq \mu_2$. Vì $(n_1 - 1)\tilde{S}_X^2 = 0,010275$; $(n_2 - 1)\tilde{S}_Y^2 = 0,0233$ nên

$$|T| = \frac{|9,363 - 9,575|}{\sqrt{\dfrac{0,010275 + 0,0233}{4 + 4 - 2}\left(\dfrac{1}{4} + \dfrac{1}{4}\right)}} = 4,01.$$

$t_{\alpha/2}(n_1 + n_2 - 2) = t_{0,025}(6) = 2,447$; $|T| > t_{\alpha/2}(n_1 + n_2 - 2)$.

Giả thuyết $\mu_1 = \mu_2$ bị bác bỏ. Ta kết luận có sự sai khác về giá trị trung bình của hai nhóm làm thí nghiệm; ít nhất một trong hai nhóm đã mắc sai số hệ thống.
#

Trường hợp này cần đặt tiếp bài toán khác, ví dụ như sau.

Tóm tắt so sánh kỳ vọng và phương sai cho ở Bảng 4.4, trong đó:

$(X_1, \ldots X_{n_1})$ tõ X víi $E[X] = \mu_1$, $V[X] = \sigma_1^2$;

$(Y_1, \ldots Y_{n_2})$ tõ Y víi $E[Y] = \mu_2$, $V[Y] = \sigma_2^2$;

$$\tilde{S}^2 = \frac{(n_1 - 1)\tilde{S}_X^2 + (n_2 - 1)\tilde{S}_Y^2}{n_1 + n_2 - 2}; \quad \overline{f} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2} = \frac{m_1 + m_2}{n_1 + n_2}.$$

Ví dụ 4.21. Hàm lượng N trong hợp chất ở Ví dụ 4.19 theo lý thuyết là $\mu_0 = 9,51(\%N)$. Hãy kiểm tra xem sai số hệ thống gây ra do nhóm I, nhóm II hay cả hai nhóm.

Ví dụ này ta đã biết cách giải (xem phần 4.5.1b). (Trả lời: Do nhóm I).

Nếu không biết μ_0, phải tiến hành loạt thí nghiệm tiếp theo rồi mới có thể đưa ra lời giải. #

d) Phương pháp so sánh từng cặp.

Giả sử (X, Y) là VTNN, hai thành phần X và Y nói chung không độc lập, $E[X] = \mu_1$; $E[Y] = \mu_2$. Ta muốn so sánh μ_1 víi μ_2.

Muốn vậy, ta lấy một mẫu $(X_1, Y_1), \ldots, (X_n, Y_n)$ về (X,Y). Ta không thể áp dụng kết quả ở phần trên được vì mặc dầu $(X_1,...,X_{n_1})$ là mẫu về X, $(Y_1,...,Y_{n_1})$ là mẫu về Y, song nói chung không độc lập.

Tuy nhiên $X_1-Y_1,...,X_n-Y_n$ lại là các BNN độc lập, chúng lập thành mẫu ngẫu nhiên từ BNN gốc $Z=X-Y$ với $E[Z]=\mu_1-\mu_2$. $\mu_1\neq\mu_2 \Leftrightarrow \mu_1-\mu_2\neq 0$, ta quay về bài toán kiểm định mẫu từ 1 BNN :

$$H_0:\mu_1-\mu_2=0\ /\ H_1:\mu_1-\mu_2\neq 0.$$

Bảng 4.4. Tóm tắt so sánh kỳ vọng và so sánh phương sai

Tham số cần KĐ: *Giả thuyết*	*Đối thuyết*	*Từ phân bố*	*Bác bỏ H_0 mức ý nghĩa α khi:* σ_1^2,σ_2^2 *đã biết (*)*	$\sigma_1^2=\sigma_2^2$ *chưa biết, mẫu nhỏ*
$\mu_1=\mu_2$	$\mu_1\neq\mu_2$	Chuẩn	$\lvert Z\rvert=\dfrac{\lvert\overline{X}-\overline{Y}\rvert}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}>z_{\alpha/2}$	$\dfrac{\lvert\overline{X}-\overline{Y}\rvert}{\tilde{S}\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}>t_{\alpha/2}(n_1+n_2-2)$
	$\mu_1>\mu_2$		$Z=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}>z_{\alpha}$	$\dfrac{\overline{X}-\overline{Y}}{\tilde{S}\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}>t_{\alpha}(n_1+n_2-2)$
	$\mu_1<\mu_2$		$Z=\dfrac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}<-z_{\alpha}$	$\dfrac{\overline{X}-\overline{Y}}{\tilde{S}\sqrt{\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}<-t_{\alpha}(n_1+n_2-2)$
$p_1=p_2$	$p_1\neq p_2$	Không - một	$\lvert Z\rvert=\dfrac{\lvert f_1-f_2\rvert}{\sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}>z_{\alpha/2}$	
	$p_1>p_2$		$Z=\dfrac{f_1-f_2}{\sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}>z_{\alpha}$	
	$p_1<p_2$		$Z=\dfrac{f_1-f_2}{\sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}<z_{1-\alpha}=-z_{\alpha}$	
			μ_1,μ_2 đã biết	μ_1,μ_2 chưa biết
$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$	Chuẩn	$\left[\begin{array}{l}F=\dfrac{S_X^{*2}}{S_Y^{*2}}<f_{1-\alpha/2}(n_1;n_2)\\ F=\dfrac{S_X^{*2}}{S_Y^{*2}}>f_{\alpha/2}(n_1;n_2)\end{array}\right.$	$\left[\begin{array}{l}F=\dfrac{\tilde{S}_X^{2}}{\tilde{S}_Y^{2}}<f_{1-\alpha/2}(n_1-1;n_2-1)\\ F=\dfrac{\tilde{S}_X^{2}}{\tilde{S}_Y^{2}}>f_{\alpha/2}(n_1-1;n_2-1)\end{array}\right.$
	$\sigma_1^2>\sigma_2^2$		$F=\dfrac{S_X^{*2}}{S_Y^{*2}}>f_{\alpha}(n_1;n_2)$	$F=\dfrac{\tilde{S}_X^{2}}{\tilde{S}_Y^{2}}>f_{\alpha}(n_1-1;n_2-1)$

(*) σ_1^2,σ_2^2 chưa biết, mẫu lớn: Không cần phân bố chuẩn, thay $\sigma_1^2(\sigma_2^2)$ bởi $S_1^2(S_2^2)$.

Ví dụ 4.22. Để khảo sát tác dụng của việc bón thêm một loại phân mới A, người ta chia mỗi thửa ruộng thí nghiệm ra làm hai mảnh: Mảnh đối chứng (không bón phân A), mảnh kia bón thêm a đơn vị phân A. Sau khi ghi lại các gia tăng năng suất Δ_i ở 15 thửa, người ta tính được độ gia tăng trung bình $\overline{\Delta}=6{,}10$

và độ lệch chuẩn 7,51. Với mức ý nghĩa 5%, hãy nhận định xem việc bón thêm phân A có tác dụng hay không? Nếu có, tìm khoảng tin cậy cho mức tăng trưởng.

Giải. Ta cần kiểm định giả thuyết

$$H_0 : \mu_2 - \mu_1 = 0 \ / \ H_1 : \mu_2 - \mu_1 > 0.$$

$T = \dfrac{6,10}{7,51}\sqrt{15} = 3,145$. Bởi vì $3,145 > 1,761 = t_{0,05}(15-1)$ nên bác bỏ giả thuyết: Việc bón thêm phân A có tác dụng nâng cao năng suất.

Tiếp theo, muốn biết việc bón thêm phân có nâng cao năng suất "đáng kể" hay không, ta cần tìm khoảng tin cậy - ví dụ 90% - cho sự gia tăng này.

$\beta = 0,9 \Rightarrow (1-\beta)/2 = 0,05;\ t_{0,05}(15-1) = 1,761$;

$\varepsilon = \dfrac{7,51}{\sqrt{15}}1,761 = 3,41$. Nhận được khoảng tin cậy $6,10 \pm 3,41$.#

Ví dụ 4.23. Hai dây chuyền sản suất những chi tiết bằng nhựa theo nguyên lý đúc - phun. Các sản phẩm được coi là hỏng nếu co ngót quá mức hoặc bị chuyển màu. Người ta lấy ra từ mỗi dây chuyền 300 sản phẩm thì thấy số sản phẩm hỏng lần lượt là 15 và 8. Kiểm tra xem hai dây chuyền này có thể xem là hoạt động như nhau hay không với mức ý nghĩa 5%; tính P - giá trị.

Giải. Gọi tỷ lệ các sản phẩm hỏng từ hai dây chuyền lần lượt là p_1, p_2. Ta xét bài toán kiểm định $H_0 : p_1 = p_2 \ / \ H_1 : p_1 \neq p_2$.

Tần suất mẫu chung là $\overline{f} = \dfrac{m_1 + m_2}{n_1 + n_2} = \dfrac{15+8}{300+300} = \dfrac{23}{600}$. Thống kê kiểm định $|Z| = |f_1 - f_2| \Big/ \sqrt{\overline{f}(1-\overline{f})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)} = 1,488$.

Tra bảng ta có $z_{\alpha/2} = z_{0,025} = 1,960$. Vì $|Z| < z_{\alpha/2}$ nên ta không bác bỏ giả thuyết, coi hai dây chuyền hoạt động như nhau.

Do $1,488 = z_{0,06837} = z_{0,1367/2}$ nên P - giá trị (2 phía) là 0,1367.

Ví dụ 4.24. Để so sánh hai phương pháp đo khi xác định natri bằng phương pháp quang phổ, người ta thu được (theo % tương đối):

Phương pháp I: $s_1 = 4,3(\%)$ với 11 bậc tự do;

Phương pháp II: $s_2 = 2,4(\%)$ với 11 bậc tự do.

Phương pháp thứ II có sai số ngẫu nhiên bé hơn hay không?

Giải. Gọi phương sai của phép đo thứ i là σ_i^2, xét bài toán kiểm định $H_0 : \sigma_1^2 = \sigma_2^2 \ / \ H_1 : \sigma_1^2 > \sigma_2^2$.

Ở đây các kỳ vọng không biết, $n_1 - 1 = n_2 - 1 = 11$, $F = 4,3^2 / 2,4^2 = 3,21$. Chúng ta chọn $\alpha = 0,05$.

Từ bảng ta không tra được $f_{0,05}(11,11)$, song lại tra được $f_{0,05}(11,10) = 2,85$; $f_{0,05}(11,12) = 2,79$. Dùng xấp xỉ tuyến tính thì $f_{0,05}(11,11) = (2,85 + 2,79)/2 = 2,82$. Vì $3,21 > 2,82$ nên bác bỏ H_0: Coi phương pháp II có độ chính xác cao hơn. #

§4.7. KIỂM ĐỊNH PHI THAM SỐ (2 tiết)

4.7.1. Kiểm định giả thuyết về luật phân bố

a) Kiểm định χ^2

Từ tổ chức đồ, hoặc một lý do nào đó có cơ sở để cho rằng hàm phân bố của X có dạng $F(x) = F(x, \theta)$ với $\theta = (\theta_1, ..., \theta_r)$ là tham số véc tơ r thành phần, ta đặt bài toán kiểm định

H_0 : Hàm phân bố của BNN X là F(x),

H_1 : Hàm phân bố của BNN X không phải là F(x).

Giả sử đã có một mẫu ngẫu nhiên $(X_1, ..., X_n)$ về X. Nếu X là BNN rời rạc với tập giá trị $\{x_1, ..., x_k\}$, từ mẫu ta tính được tần số của các giá trị này là $n_1, ..., n_k$. Khi giả thuyết H_0 là đúng thì có thể tính được các xác suất

$$p_i = P\{X = x_i\} = p_i(\theta_1, ..., \theta_r), \quad i = 1, ..., k .$$

Nếu X là BNN liên tục với hàm mật độ $f(x, \theta) = f(x, \theta_1, ..., \theta_r)$, ta hãy chia tập giá trị của X làm k khoảng rời nhau $(a_0; a_1]$; $(a_1; a_2]$; ...; $(a_{k-1}; a_k]$. Gọi n_i là số lần các trị số X_j rơi vào khoảng $(a_{i-1}; a_i]$, còn gọi là tần suất mẫu. Rõ ràng $n_1 + ... + n_k = n$. Ta nhận được bảng phân bố ghép lớp sau đây

X_i	$(a_0; a_1]$	$(a_1; a_2]$	. . .	$(a_{k-1}; a_k]$
n_i	n_1	n_2	. . .	n_k

Khi giả thuyết là đúng, ta tính được các xác suất

$$p_i = P\{a_{i-1} < X \le a_i\} = p_i(\theta_1, ..., \theta_r), \quad i = 1, ..., k .$$

Trong cả hai trường hợp gián đoạn và liên tục ta đều tính được

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i} .$$

Chứng minh được rằng, với một số điều kiện có thể thoả mãn trong những tình huống thông thường (xem [2], tr 239), không phụ thuộc vào dạng phân bố của X, khi $n \to \infty$, χ^2 như trên có phân bố xấp xỉ phân bố khi bình phương với $k - r - 1$ bậc tự do.

Lưu ý: i) r là số tham số; ví dụ, với phân bố chuẩn thì $r = 2$, phân bố chuẩn khi phương sai (hay kỳ vọng) đã biết thì $r = 1$, phân bố Poisson thì $r = 1$… Khi F(x) là hàm phân bố cụ thể nào đó thì xem như không có tham số, hay số tham số $r = 0$, đưa ra bởi Pearson.

ii) Khi tính p_i ta phải dùng ƯL của các tham số $\theta_1, ..., \theta_r$. Chúng có thể được ƯL theo phương pháp hợp lý cực đại.

iii) Các khoảng chia được chọn sao cho các tần số n_i không quá bé - trừ khoảng đầu và khoảng cuối - thường thì $n_i \ge 5$, tốt ra thì $n_i \ge 10$. Số các khoảng cũng không quá bé $(k \ge 5)$.

Từ đó, với mức ý nghĩa α, giả thuyết bị bác bỏ nếu

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i} > \chi^2_\alpha(k - r - 1) . \qquad (4.7.1)$$

Thủ tục kiểm định trên được gọi là kiểm định khi bình phương.

Công thức (4.7.1) được hiệu chỉnh tốt hơn nếu ta thay n_i bởi $n_i - 0,5$. Hiệu chỉnh này có tác dụng khi bậc tự do là nhỏ; đặc biệt, khi bậc tự do bằng 1 thì bắt buộc phải dùng hiệu chỉnh này.

Ví dụ 4.25. Số lần hỏng hóc của một quá trình sản xuất ứng với 8 kíp làm việc được ghi lại thành bảng

Kíp i	1	2	3	4	5	6	7	8	Σ
Số lần hỏng n_i	16	18	19	16	24	19	17	15	144

Số lần hỏng xảy ra ở các kíp có thể coi là như nhau hay không?

Giải. Đặt $p_i = n_i / n$ là tỷ lệ hỏng hóc do kíp i gây ra, $i = 1,...,8$. Gọi X là chỉ số của kíp xảy ra hỏng hóc, X có bảng xác suất

X	1	2	3	.	.	.	8
P	p_1	p_2	p_3	.	.	.	p_8

Ta phải kiểm tra giả thuyết X có phân bố đều trên tập $\{1;...;8\}$, ($\Leftrightarrow p_1 = p_2 = ... = p_8 = 1/8$) với đối thuyết X không có phân bố đều.

Ở đây r = 0. Bởi vì $p_i = 1/8$ ta tính được

$$\chi^2 = \sum_{i=1}^{n} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{8} (n_i - 18)^2 / 18 = 3,111.$$

Mặt khác, $\chi^2_\alpha(k-1) = \chi^2_{0,05}(8-1) = 14,07$. Vậy $\chi^2 < \chi^2_\alpha(k-1)$. (P-giá trị trong trường hợp này là 0,875, vượt xa ngưỡng 0,05), ta không bác bỏ H_0: Coi số lần hỏng hóc ở các kíp là như nhau. #

Ví dụ 4.26. Trong một cuộc tổng điều tra dân số, người ta biết tỷ lệ phân bố dân cư theo dân tộc ở một vùng núi như sau

Dân tộc	Kinh	Tày	Thái	Mường	Khác	Cộng
Tỷ lệ (%)	21	15	30	17	17	100

Sau 5 năm, điều tra ngẫu nhiên 600 người ở vùng đó thu được

Dân tộc	Kinh	Tày	Thái	Mường	Khác	Cộng
n_i	140	100	140	100	120	600

Xét xem cơ cấu dân cư ở đó đã có thay đổi đáng kể hay chưa?

Ta phải kiểm tra xem cơ cấu dân cư có còn tuân theo bảng phân bố (xác suất) đã cho ở trên hay không. Như vậy r = 0. Ta có

$$\chi^2 = \frac{(140 - 600.0,21)^2}{600.0,21} + ... + \frac{(120 - 600.0,17)^2}{600.0,17} = 14.67$$

$$\chi^2 > \chi^2_\alpha(k-1) = \chi^2_{0.05}(5-1) = 9,488.$$

Từ đó ta bác bỏ giả thuyết, coi cơ cấu dân cư đã thay đổi.

Ví dụ 4.27. Số sản phẩm tạo ra được từ 200 chu kỳ sản suất được ghi lại thành bảng

Số SP	10-13 13-16 16-19 19-22 22-25 25-28 28-31 31-34 34-37
Số lần	12 20 23 40 32 21 22 20 10

Với mức ý nghĩa $\alpha = 0,05$ có thể coi năng suất trong một chu kỳ tuân theo luật chuẩn được không? (Xem [1])

b) Chọn khoảng để tần số lý thuyết đều nhau.

Ví dụ 4.28. Để kiểm tra xem điện áp đầu vào X của một loại máy tính xách tay có tuân theo luật chuẩn hay không, người ta thử nghiệm 100 lần đo và thu được điện áp trung bình $\overline{x} = 5,04\text{V}$ với độ lệch chuẩn $s = 0,08\text{V}$.

Giả sử muốn chia tập giá trị của X thành $k = 8$ khoảng với xác suất đều nhau trên mỗi khoảng. Trước hết tần số lý thuyết ứng với mỗi khoảng sẽ là $n_0 = n.p_i = 100/8 = 12,5$. Sử dụng hàm phân bố chuẩn tắc dễ tính được các khoảng giá trị cho BNN chuẩn tắc là $[0; 0,320)$, $[0,320; 0,675)$, $[0,675; 1,150)$; $[1,150; +\infty)$ và các khoảng đối xứng với chúng qua gốc toạ độ. Từ đó suy ra cách chia tập giá trị của X (từ điều kiện $\dfrac{x-\overline{x}}{s} = \dfrac{x-5,04}{0,08}$ nằm trên các khoảng này). Tính toán tần số quan sát người ta ghi kết quả vào Bảng 4.6. Giá trị của thống kê kiểm định là

$$\chi^2 = \frac{1}{12,5}\left[(12-12,5)^2 + (14-12,5)^2 + ... + (14-12,5)^2\right] = 8.$$

Lại có $\chi^2_\alpha(k-2-1) = \chi_{0,05}{}^2(5) = 11,07$. Vì $\chi^2 < \chi^2_\alpha$ nên không bác bỏ giả thuyết, coi X có phân bố chuẩn.

Bảng 4.6. Số liệu điện áp

Lớp	Tần số quan sát	Tần số lý thuyết
$x < 4,948$	12	12,5
$4,948 \le x < 4,986$	14	12,5
$4,986 \le x < 5,014$	12	12,5
$5,014 \le x < 5,040$	14	12,5
$5,040 \le x < 5,066$	12	12,5
$5,066 \le x < 5,094$	11	12,5
$5,094 \le x < 5,132$	12	12,5
$5,132 \le x$	13	12,5
Tổng	100	100

c) Tiêu chuẩn Kolmogorov-Smirnoff

4.7.2. Kiểm định về tính độc lập của hai biến ngẫu nhiên

Giả sử ta có mẫu ngẫu nhiên $(X_1, Y_1),...,(X_n, Y_n)$ từ VTNN gốc (X, Y). Dựa vào mẫu ngẫu nhiên này ta muốn kiểm tra xem hai thành phần X, Y có độc lập với nhau hay không.

a) Trường hợp liên tục. Nếu (X, Y) liên tục, ta chia tập giá trị của thành phần thứ nhất thành r khoảng rời nhau $(a_0;a_1];\ (a_1;a_2];\ ...;(a_{r-1};a_r]$, của thành phần thứ hai thành s khoảng rời nhau $(b_0;b_1];(b_1;b_2];...;(b_{s-1};b_s]$. Gọi n_{ij} là số quan sát từ mẫu mà thành phần thứ nhất rơi vào khoảng thứ i: $(a_{i-1};a_i]$, thành phần thứ hai rơi vào khoảng thứ j: $(b_{j-1};b_j]$, $i=1,...,r;\ j=1,...,s$. Kết quả được sắp xếp thành bảng sau, gọi là bảng phân bố ghép lớp

Y \ X	$(b_0;b_1]$	$(b_1;b_2]$	. . .	$(b_{s-1};b_s]$	Σ
$(a_0;a_1]$	n_{11}	n_{12}	. . .	n_{1s}	m_1
$(a_1;a_2]$	n_{11}	n_{12}	. . .	n_{1s}	m_2
. . .	. . .	. . .	. . .	. .	. .
$(a_{r-1};a_r]$	n_{r1}	n_{r2}	. . .	n_{rs}	m_r
Σ	n_1	n_2	. . .	n_s	n

Như vậy, số giá trị quan sát được là $k=rs$ các giá trị n_{ij}.

Đặt $p_i = P(X\in(a_{i-1};a_i])$, $q_j = P(Y\in(b_{j-1};b_j])$. Giả thuyết về tính độc lập của X và Y xem như tương đương với $p_{ij}=p_i q_j, \forall i, j$. Vậy chỉ cần ƯL r + s tham số này là đủ. Tuy nhiên, lại có hai hệ thức ràng buộc là $\sum_{i=1}^{r} p_i = \sum_{j=1}^{s} q_j = 1$, nên số tam số cần ƯL chỉ còn $r+s-2$. Các ƯL của xác suất biên tương ứng là $\hat{p}_i = m_i/n;\ \hat{q}_j = n_j/n$.

Theo (4.7.1), thống kê kiểm định được chọn là

$$\chi^2 = \sum_{i,j}\left(n_{ij} - n\frac{m_i}{n}\frac{n_j}{n}\right)^2 \Big/ \left(n\frac{m_i}{n}\frac{n_j}{n}\right) \quad \text{hay}$$

$$\chi^2 = n\sum_{i,j}\left(n_{ij} - \frac{m_i n_j}{n}\right)^2 \Big/ (m_i n_j) = n\left(\sum_{i,j}\frac{n_{ij}^2}{m_i n_j} - 1\right). \quad (4.7.6)$$

Nếu giả thuyết là đúng thì với n lớn, χ^2 có phân bố xấp xỉ phân bố khi bình phương với $rs-(r+s-2)-1=(r-1)(s-1)$ bậc tự do. Tương tự các mục trên, giả thuyết bị bác bỏ nếu

$$\chi^2 > \chi^2_\alpha((r-1)(s-1)). \quad (4.7.7)$$

Lưu ý. Nếu thấy $n_{ij}<5$ thì ta ghép các khoảng gần nhau lại để $n_{ij}\ge 5\ \forall i, j$.

b) Trường hợp rời rạc

Nếu (X,Y) là VTNN rời rạc thì khi thay các khoảng bởi các giá trị mà từng thành phần có thể nhận, ta vẫn nhận được bảng như trên, gọi là bảng phân bố tần số. Phần còn lại của kiểm định được tiến hành tương tự.

Trường hợp một hoặc cả hai thành phần X, Y có thể nhận quá nhiều giá trị, ta xem như đó là (các) BNN liên tục và nên ghép những giá trị gần nhau thành một lớp, ta lại trở về trường hợp liên tục.

c) Kiểm định về tính độc lập giữa hai dấu hiệu định tính

Phương pháp trên cũng áp dụng rất thành công để kiểm tra sự độc lập của hai dấu hiệu định tính. Giả sử mỗi cá thể trong tổng thể mang 2 dấu hiệu A và B, ta cần nghiên cứu hai dấu hiệu A và B này. Lại giả sử có thể chia dấu hiệu A, B tương ứng thành r và s dấu hiệu thành phần: $A = \{A_1,...,A_r\}$; $B = \{B_1,...,B_s\}$. Nếu có cơ sở để đưa ra giả thuyết các dấu hiệu A và B độc lập, ta xét bài toán kiểm định

$$H_0: \text{ A và B độc lập } / \ H_1: \text{ A và B phụ thuộc.}$$

Để kiểm định, ta lập mẫu kích thước n và trình bày kết quả thành bảng với n_{ij} là tần số ứng với dấu hiệu A_i, B_j đồng thời:

A \ B	B_1 B_2 . . . B_s	Σ
A_1	n_{11} n_{12} . . . n_{1s}	m_1
A_2	n_{21} n_{22} . . . n_{2s}	m_2
. . .	. . . .	. .
A_r	n_{r1} n_{r2} . . . n_{rs}	m_r
Σ	n_1 n_2 . . . n_s	n

Để giải bài toán này ta xét biến X và Y như sau. Nếu cá thể trong tổng thể mang dấu hiệu A_i thì ta đặt $X = i$, nếu mang dấu hiệu B_j thì ta đặt $Y = j$. Véc tơ (X,Y) biến đổi giá trị từ cá thể nọ sang các thể kia, xem như là VTNN xác định trong tổng thể. Ta quay về trường hợp VTNN rời rạc.

Ví dụ 4.29. Ban giám đốc một ngân hàng muốn kiểm tra xem giá trị của các quyển sổ gửi tiền tiết kiệm có độc lập với khu vực đặt các chi nhánh phát hành sổ hay không. Một cuộc khảo sát được tiến hành trên 900 cuốn sổ. Sổ được coi là nhỏ nếu có số dư không quá 10; sổ được coi là thông thường nếu số dư từ 10 đến 100; sổ được coi là quan trọng nếu số dư trên 100 (đơn vị: triệu VNĐ). Sau đây là bảng kết quả.

Loại sổ \ Địa chỉ	Nhỏ	Thường	Q. trọng	Tổng
Hà Nội	27	189	54	270
TP H. C. M.	33	124	23	180
Các nơi khác	103	309	38	450
Tổng	163	622	115	900

Ta cần kiểm tra giả thuyết số dư của các cuốn sổ là độc lập với các nơi gửi các cuốn sổ đó. Từ bảng ta tính được

$$\chi^2 = n\left(\sum_{i,j}\frac{n_{ij}^2}{n_i\, m_j} - 1\right) = 900\left[\frac{27^2}{163.270} + ... + \frac{38^2}{115.450} - 1\right] = 33{,}388\,.$$

Vì $\chi^2 > 9{,}49 = \chi^2_{0.05}((3-1)(3-1))$ nên ta bác bỏ giả thuyết; xem rằng có sự phụ thuộc giữa địa chỉ đặt các chi nhánh với giá trị các cuốn sổ tiết kiệm.

BÀI TẬP: Kiểm định dùng một mẫu (1 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 13: Kiểm định giả thuyết và ước lượng tham số (tiếp)

Chương, mục: 4

Tiết thứ: 49-52 Tuần thứ: 13

Mục đích, yêu cầu:

- Nắm Cách sử dụng hệ số tương quan trong kỹ thuật

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

Hệ số tương quan

Bài tập về kiểm định giả thuyết

HỆ SỐ TƯƠNG QUAN MẪU

e) Hệ số tương quan mẫu

Hệ số tương quan mẫu của mẫu ngẫu nhiên (X_1, Y_1), ..., (X_n, Y_n) từ VTNN gốc (X, Y), ký hiệu r_{XY}, cho bởi công thức

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}. \qquad (5.1)$$

Giá trị cụ thể của nó ứng với mẫu cụ thể $(x_1, y_1), ..., (x_n, y_n)$ cho bởi

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}. \qquad (5.2)$$

Hệ số tương quan mẫu r_{XY} thể hiện mối quan hệ tuyến tính giữa X và Y. Đó là một ƯL (xấp xỉ) cho hệ số tương quan ρ_{XY}.

Rất tiếc, r_{XY} lại không phải là một ƯL tốt của ρ_{XY}: Nó là ƯL chệch! Thành thử, (5.1), (5.2) chỉ được dùng làm ƯL khi n lớn, chẳng hạn $n \geq 100$.

Để có ƯL tốt hơn, người ta xét một hàm của r_{XY}, đó là:

$$Z = \frac{1}{2}\ln\frac{1 + r_{XY}}{1 - r_{XY}}.$$

Với n đủ lớn, chẳng hạn $n \geq 50$, chứng minh được Z xấp xỉ phân bố chuẩn

$$N\left(\frac{1}{2}\ln\frac{1 + \rho_{XY}}{1 - \rho_{XY}} + \frac{\rho_{XY}}{2(n-1)};\ \frac{1}{n-3}\right). \qquad (5.3)$$

Từ đó tìm ra được khoảng tin cậy xấp xỉ cho Z; dùng hàm ngược tính ngược lại, thu được khoảng tin cậy xấp xỉ cho ρ_{XY}.

Xấp xỉ (5.3) cũng dùng để kiểm định về hệ số tương quan. Chẳng hạn, xét kiểm định 2 phía

$$H_0 : \rho = \rho_0 / \ \rho \neq \rho_0,$$

Giả thuyết bị bác bỏ ở mức α nếu

$$\left|\frac{Z - E[Z]}{\sqrt{V[z]}}\right| = \left|(Z - E[Z])\sqrt{n-3}\right| > z_{\alpha/2},$$

trong đó $E[Z] = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2(n-1)}$.

BÀI TẬP: Kiểm định dùng một mẫu (tiếp - 1 tiết)
Kiểm định dùng hai mẫu (2 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	**Bài tập chuẩn bị cho cả chương 5** Tài liệu [1]: 5(1- 4 - <u>5</u>- <u>6</u>- <u>8</u> - <u>9</u>- <u>12</u>- 14 - <u>15</u>) Tài liệu [2]: Tr187-189: 3, 4, 6, <u>8</u>, <u>10</u>, 14, <u>16</u>, <u>17</u>, 22, 26, 28
Tài liệu	Tài liệu [1], tr ....

Bài giảng 14: Mô hình hồi quy tuyến tính

Chương, mục: 5

Tiết thứ: 53-56 Tuần thứ: 14

Mục đích, yêu cầu:

- Nắm được cách lập mô hình, một số kiểm định, một số cách sử dụng của MH TT đơn, bội

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

Chương 5. Mô hình hồi quy tuyến tính

§5.1. Mô hình hồi quy tuyến tính đơn

§5.2. Mô hình hồi quy tuyến tính bội

Chương 5

MÔ HÌNH HỒI QUY TUYẾN TÍNH

§ 5.1. MÔ HÌNH HỒI QUY TUYẾN TÍNH ĐƠN (1 tiết)

5.1.1. Vấn đề mô hình hồi quy

Nhiều bài toán trong khoa học kỹ thuật đòi hỏi khảo sát quan hệ giữa hai hoặc nhiều biến. Lấy làm ví dụ, chúng ta xét số liệu ở Bảng 5.1, ở đó y chỉ thị độ sạch của oxy sinh ra trong quá trình chưng cất hóa học, còn x là nồng độ phần trăm của hydrocarbon có mặt ở bình ngưng bộ phận chưng cất.

Bảng 5.1. Độ sạch của oxy ứng với tỷ lệ phần trăm hydrocarbon

TT	x(%)	y(%)	TT	x(%)	y(%)	TT	x(%)	y(%)
1	0.99	90.01	8	1.23	91.77	15	1.11	89.85
2	1.02	89.05	9	1.55	99.42	16	1.2	90.39
3	1.15	91.43	10	1.4	93.65	17	1.26	93.25
4	1.29	93.74	11	1.19	93.54	18	1.32	93.41
5	1.46	96.73	12	1.15	92.52	19	1.43	94.98
6	1.36	94.45	13	0.98	90.56	20	0.95	87.33
7	0.87	87.59	14	1.01	89.54	21	1.32	94.01

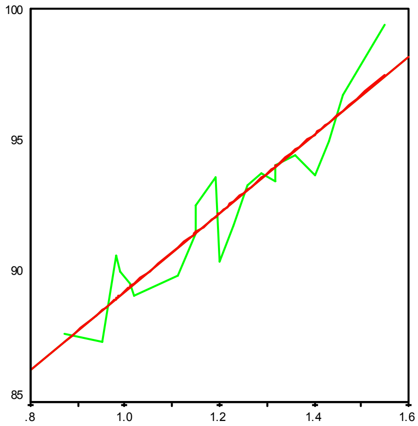
Để tổng quát hóa, chúng ta nên dùng mô hình xác suất bằng cách coi Y là BNN mà ứng với giá trị x của biến X thì

$$Y = f(x) + \varepsilon \qquad (5.1.2)$$

với ε là sai lầm ngẫu nhiên.

Trước hết chúng ta xét trường hợp đơn giản nhất, cũng rất hay xảy ra trong thực tế, khi $f(x) = ax + b$. Khi đó (5.1.2) trở thành

$$Y = ax + b + \varepsilon \qquad (5.1.3)$$



Hình 5.1. Đồ thị rải điểm, đường hồi quy cho số liệu độ sạch của oxy

Mô hình (5.1.3) được gọi là mô hình hồi quy (MHHQ) tuyến tính đơn; x được gọi là biến hồi quy (hay biến độc lập, biến giải thích), Y được gọi là biến phản hồi (hay biến phụ thuộc, biến được giải thích); a, b được gọi là các tham số hồi quy, a: hệ số chặn, b: hệ số góc; đường thẳng $y = ax + b$ được gọi là đường hồi quy (lý thuyết).

Mô hình được gọi là tuyến tính vì nó tuyến tính với các tham số a, b (a, b có lũy thừa 1); được gọi là đơn vì có một biến hồi quy. Ở bài §5.2 chúng ta sẽ xét mô hình hồi quy bội với ít nhất 2 biến hồi quy. Người ta cũng xét mô hình hồi quy phi tuyến, ở đó hàm hồi quy là hàm phi tuyến của các tham số (xem [1], [9]).

Giả sử ở quan sát thứ i biến X nhận giá trị x_i, biến Y nhận giá trị y_i và sai lầm ngẫu nhiên là ε_i. Như vậy, dưới dạng quan sát, mô hình (5.1.3) trở thành

$$\begin{cases} y_1 = a + bx_1 + \varepsilon_1 \\ . \quad . \quad . \quad . \quad . \quad . \quad . \\ y_n = a + bx_n + \varepsilon_n \end{cases} \qquad (5.1.4)$$

Lưu ý rằng y_i là các BNN.

5.1.2. Ước lượng hệ số hồi quy

Bây giờ giả sử các BNN $y_1, ..., y_n$ nhận các giá trị cụ thể nào đó, vẫn ký hiệu là $y_1, ..., y_n$. Khi đó

$$\varepsilon_i = y_i - (ax_i + b)$$

(5.1.5)

thể hiện độ lệch của quan sát thứ i so với đường hồi quy lý thuyết (xem Hình 5.2). Tổng bình phương các độ lệch

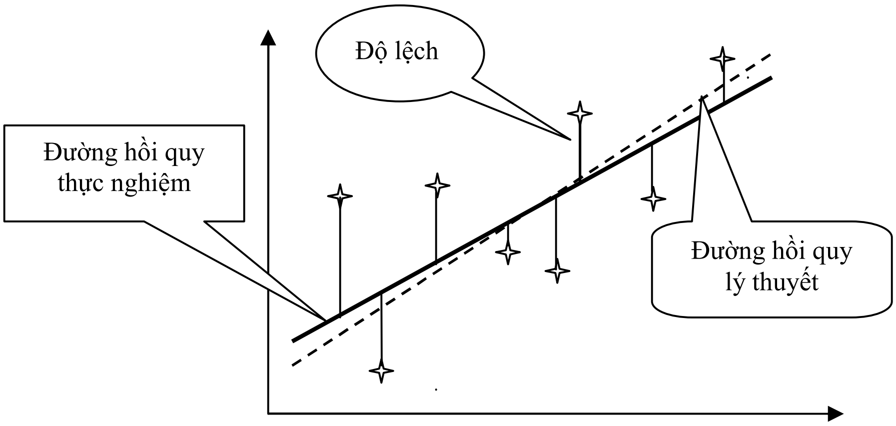
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

thể hiện "chất lượng" của việc xấp xỉ số liệu bởi đường hồi quy lý thuyết. Ta không thể biết đường hồi quy lý thuyết, việc ta có thể làm là tìm các hệ số a, b để

$$\ell(a,b) = \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \to \min . \qquad (5.1.6)$$

Vì $\ell(a,b)$ là đa thức bậc 2 của 2 ẩn a, b; điều kiện cần để nó đạt cực tiểu là

$$\frac{\partial \ell}{\partial a} = \frac{\partial \ell}{\partial b} = 0 . \qquad (5.1.7)$$



Hình 5.2. Độ lệch và các đường hồi quy lý thuyết, thực nghiệm

Thực ra chứng minh được đây cũng là điều kiện đủ. Đây là hệ 2 phương trình tuyến tính bậc nhất của a, b. không khó khăn gì ta tính được nghiệm của hệ này là:

$$\begin{cases} \hat{b} = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{S_{XX}/n} \\ \hat{a} = \overline{y} - \hat{b}\overline{x} \end{cases} \qquad (5.1.8)$$

trong đó

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ; \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i ; \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i ; \quad S_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2 . (5.1.9)$$

Với các ƯL này ta được phương trình hồi quy thực nghiệm

$$y = \hat{a}x + \hat{b} . \qquad (5.1.10)$$

5.1.3. Tính chất của ước lượng của các hệ số hồi quy

Từ (5.8) ta có ngay $\overline{y} = \hat{a} + \hat{b}\overline{x}$. Như vậy, đường hồi quy đi qua điểm "trung tâm" $(\overline{x}, \overline{y})$ của số liệu.

Giả thiết:

$\varepsilon_1, ..., \varepsilon_n$ độc lập, cùng phân bố chuẩn $N(0;\sigma^2)$. (5.1.11)

Khi đó ƯL hệ số có những tính chất thống kê tốt thể hiện ở định lý sau.

Định lý 5.1. Khi điều kiện (5.1.11) thỏa mãn thì:

i) $\hat{a}$ và $\hat{b}$ lần lượt là ƯL không chệch của tham số a và b:

$$E[\hat{a}] = a; \qquad E[\hat{b}] = b \qquad (5.1.12)$$

ii) Phương sai của các ƯL $\hat{a}$ và $\hat{b}$ được tính như sau

$$\sigma_a^2 = V[\hat{a}] = \sigma^2 \left(\frac{1}{n} + \frac{(\overline{x})^2}{S_{XX}} \right),$$

$$\sigma_{b}^{2} = V[\hat{b}] = \frac{\sigma^2}{S_{XX}} \qquad (5.1.13)$$

iii) ƯL không chệch của phương sai chung σ^2 của mô hình cho bởi

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (5.1.14)$$

với

$\hat{y}_i = \hat{a} + \hat{b}x_i$: dự báo của quan sát thứ i

$e_i = y_i - \hat{y}_i$: phần dư thứ i.

Định nghĩa. Đối với mô hình HQTT đơn, sai số chuẩn hóa (thực nghiệm) của hệ số góc và hệ số chặn lần lượt được xác định bởi

$$se(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}}; \qquad se(\hat{a}) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]} \qquad (5.1.15)$$

trong đó, $\hat{\sigma}^2$ được tính theo (5.1.14).

5.1.4. Kiểm định giả thuyết

a) Sử dụng kiểm định T

Hệ số góc là tham số quan trọng nhất của MHHQ tuyến tính đơn. Xét bài toán kiểm định giả thuyết hai phía:

$$H_0: \ b = b_0 \,/\, H_1: \ b \neq b_0. \qquad (5.1.16)$$

Bác bỏ H_0 (ở mức ý nghĩa α) nếu

$$|T_b| = \frac{|\hat{b} - b_0|}{se(\hat{b})} = \frac{|\hat{b} - b_0|}{\sqrt{\hat{\sigma}^2 / S_{XX}}} > t_{\frac{\alpha}{2}}(n-2). \qquad (5.1.18)$$

Trường hợp đặc biệt quan trọng là khi $b_0 = 0$:

$$H_0: \ b = 0 \,/\, H_1: \ b \neq 0. \qquad (5.1.19)$$

Điều này liên quan đến ý nghĩa (hay tác dụng) của hồi quy (significance of regression): Nếu không bác bỏ H_0 (coi $b = 0$) thì có nghĩa rằng không có một quan hệ tuyến tính nào giữa X và Y (có thể là quan hệ thực sự của X và Y là quan hệ phi tuyến), sự thay đổi của biến X không kéo theo sự thay đổi dự đoán biến Y, X không có (hoặc rất ít) tác dụng để dự đoán Y; dự đoán cho Y tốt nhất nên dùng $\overline{Y}$.

Tương tự, giả thuyết liên quan đến hệ số chặn là

$$H_0: \ a = a_0 \,/\, H_1: \ a \neq a_0. \qquad (5.1.20)$$

Giả thuyết bị bác bỏ ở mức α nếu

$$|T_a| = \frac{|\hat{a} - a_0|}{se(\hat{a})} = \frac{|\hat{a} - a_0|}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]}} > t_{\frac{\alpha}{2}}(n-2). \qquad (5.1.22)$$

5.1.5. Khoảng tin cậy

a) Khoảng tin cậy của các tham số

b) Khoảng tin cậy cho đáp ứng trung bình

$$\begin{cases} \omega = t_{\alpha/2}(n-2)\sqrt{\hat{\sigma}^2\left[\dfrac{1}{n}+\dfrac{(x_0-\overline{x})^2}{S_{XX}}\right]}, \\ \hat{y}_0 = \hat{a}+\hat{b}x_0. \end{cases} \qquad (5.1.27)$$

c) Dự đoán quan sát tương lai

$$\hat{y}_0 = \hat{a}+\hat{b}x_0. \qquad (5.1.28)$$

Ví dụ 5.1. Thông thường, người ta vẫn nghĩ mức tiêu thụ nhiên liệu không phụ thuộc vào việc lái xe nhanh hay chậm. Để kiểm tra người ta cho chạy thử một chiếc xe con ở nhiều vận tốc khác nhau từ 45 đến 70 dặm/giờ. Kết quả ghi thành bảng

Vận tốc	45	50	55	60	65	70	75
Mức tiêu thụ (ml/gal)	24,2	25,0	23,3	22,0	21,5	20,6	19,8

Liệu có thể thay đổi cách nghĩ rằng mức tiêu thụ nhiên liệu không phụ thuộc vào vận tốc xe? Tìm các khoảng tin cậy 95% cho giá trị trung bình và của quan sát tương lai của mức tiêu thụ nhiên liệu khi xe ở vận tốc 50 ml/h .

Giải. Chúng ta xét mô hình HQTT đơn $Y = a+bx+\varepsilon$, trong đó Y là mức tiêu thụ nhiên liệu, x là vận tốc xe. Cần phải xét xem hệ số b có bằng không hay không. Muốn thế ta xét bài toán kiểm định:

$$H_0: \ b=0 / H_1: \ b\neq 0.$$

Tính toán các thống kê liên quan ta được

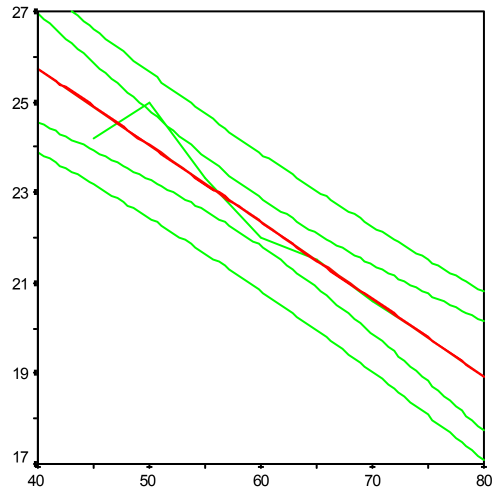
$\overline{x}=60;\quad S_{XX}=700;\ \overline{y}=22,343;\ S_{YY}=21,757;\ S_{XY}=-119$

$\hat{a}=32,543;\qquad \hat{b}=-0.17;\qquad SS_R=1.527$

Mô hình thực nghiệm: $y=32,54-0,17x$.

Tra bảng ta thấy $t_{0.025}(5)=2,571$. Theo (5.1.26), khoảng tin cậy 95% của b là $(-0,170\pm 2.571\sqrt{\dfrac{1.527}{3500}})=(-0,224;-0,116)$. Khoảng này không chứa điểm 0, vậy ta bác bỏ giả thuyết $b=0$ với mức ý nghĩa 5%; coi $b\neq 0$, tức là mức tiêu thụ nhiên liệu phụ thuộc vào vận tốc xe. Cũng có thể tính trực tiếp để bác bỏ $b=0$:

$$T_b=\frac{|\hat{b}-b_0|}{\sqrt{\hat{\sigma}^2/S_{XX}}}=\frac{|-0,17|}{\sqrt{\dfrac{0,305426}{700}}}=8,13>2,571=t_{0.025}(5).$$



Hình 5.3. Khoảng tin cậy (2 đường Hyperbol giữa)
và khoảng dự đoán (2 đường hyperbol ngoài) cho mức tiêu thụ nhiên liệu

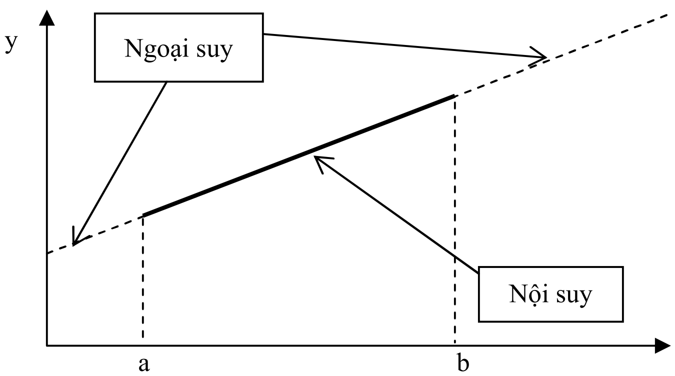
Dùng (5.1.27) và (5.1.29), khoảng tin cậy và khoảng dự đoán 95% tại vận tốc 50ml/h là

$$\left(24,04 \pm 2,571\left[\sqrt{\frac{1}{7}+\frac{(50-60)^2}{700}}\right]\right) = (24,04 \pm 1,37) = (22,67;\ 24,41)$$

$$\left(24,04 \pm 2,571\sqrt{1+\frac{1}{7}+\frac{(50-60)^2}{700}}\right) = (24,04 \pm 2,92) = (21,12;\ 26,96)$$

d) Lưu ý khi sử dụng MHHQ

- *Trường hợp nội suy.*
- *Trường hợp ngoại suy.*



5.1.6. Tí

a) Phân

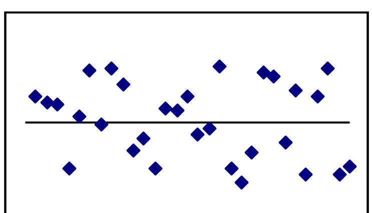
Phần dư $e_i = y_i - \hat{y}_i$

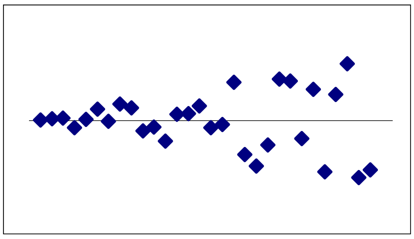
Phần dư chuẩn hóa $d_i = e_i / \sqrt{\hat{\sigma}^2}$, $i = 1,...,n$.

Có khoảng 95% các phần dư chuẩn hóa rơi vào khoảng (-2; 2)

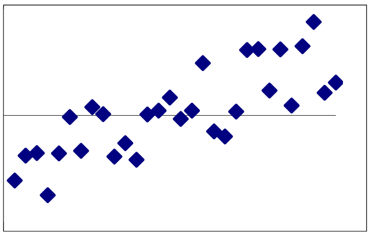
Hơn nữa, đồ thị d_i phải có dạng bình thường, tập trung "đều đặn" trong dải (-2; 2) quanh trục hoành như dạng (a) ở Hình 5.5. Vi phạm điều đó, chẳng hạn nếu nó có dạng (b), (c), (d) thì phải sửa chữa mô hình, hay tìm mô hình khác và phân tích lại.

(a) (b)

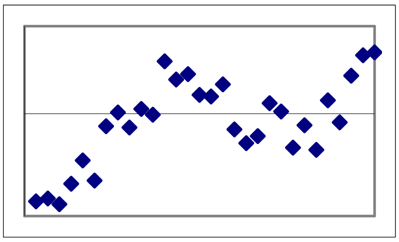




(c)



(d)



Hình 5.5. Dáng điệu phần dư

b) Hệ số xác định (coefficient of determination)

Hệ số xác định ký hiệu bởi R^2 được tính theo công thức sau:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}. \qquad (5.1.30)$$

Theo (5.1.23'), tính chất của hệ số xác định là

$$0 \le R^2 \le 1.$$

Gọi r_{XY} là hệ số tương quan mẫu của các cặp điểm (x_i, y_i) (xem mục 4.1.2e) thì ta có thể thấy

$$R^2 = r_{XY}^2. \qquad (5.1.30')$$

Giá trị R^2 thường được xem như một chỉ thị cho tính "tốt" của mô hình: Khi giá trị này gần bằng 1, mô hình phù hợp tốt; khi giá trị này nhỏ, gần bằng 0, mô hình không phù hợp với số liệu, cần tìm mô hình khác. Tuy nhiên, cần thận trọng, ngưỡng nào cho một mô hình cụ thể lại là điều ta chưa biết, ít ra là đến thời điểm này.

Lưu ý. Liên quan đến máy tính bỏ túi CASIO, ta có thể tính $\hat{\sigma}^2$ như sau:

$$R^2 = 1 - \frac{n-2}{n}\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1 - \frac{n-2}{n}\frac{\hat{\sigma}^2}{(y\sigma n)^2}$$

$$\Rightarrow \hat{\sigma}^2 = \frac{n}{n-2}(1 - R^2)(y\sigma n)^2 \qquad (5.1.31)$$

với $\quad (y\sigma n)^2 = S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$

Ví dụ 5.2. Trong nhà máy sản xuất các linh kiện bán dẫn, linh kiện hoàn chỉnh là dây được bó xếp lại thành một cái khung. Người ta quan tâm đến 3 biến: lực kéo (số đo của lực làm cho khung bị hỏng), độ dài của dây, và chiều cao của khuôn đúc. Số liệu có 25 quan sát thể hiện ở 4 cột đầu Bảng 5.5.

Trước hết ta quan tâm đến mối quan hệ giữa lực kéo y và độ dài x_1 của dây, ở đây để tiện ta vẫn ký hiệu là x. Thể hiện số liệu lên đồ thị, dường như đây là quan hệ tuyến tính. Chúng ta dùng mô hình $Y = ax + b + \varepsilon$ để lọc số liệu. Ta tính được:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 8,24; \qquad S_{XX} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = 698,56;$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 29,0328; \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i = 320,3388;$$

$$(\sigma yn)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = 224,237.$$

Từ đó ƯL của các hệ số là

$$\hat{b} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{S_{XX}/n} = 2,9027; \quad \hat{a} = \overline{y} - \hat{b}\overline{x} = 5,115.$$

Ta thu được phương trình :

$$Y = 5,115 + 2,9027x. \qquad (5.1.32)$$

ƯL của σ^2 có thể tính theo $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Tuy nhiên trước hết ta tìm hệ số xác định:

$$R^2 = \frac{SS_R}{SS_T} = \left(\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2\right) / \left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right) = 0.964.$$

Đây là giá trị khá lớn. Ta nói có 96,4% số liệu được giải thích bởi mô hình. Theo (5.1.31) thì

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{n}{n-2}(1 - R^2)(y\sigma n)^2 = 9,5696 = 3,0934^2.$$

Bây giờ ta kiểm định hệ số $b = 0$. Theo (5.1.15),

$$se(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} = 0.1179 \Rightarrow T_b = \frac{|\hat{b} - 0|}{se(\hat{b})} = \frac{2,9027}{0,1179} = 24,80.$$

P – giá trị của phân bố Student 23 bậc tự do ứng với giá trị 24,80 là 0,000. Vậy ta chấp nhận giả thuyết $b \neq 0$.

Bây giờ ta xét phân tích phương sai.

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = 5885,9 \Rightarrow SS_R / 1 = 5885,9,$$

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 220,1 \Rightarrow \hat{\sigma}^2 = \frac{SS_E}{n-2} = 9,569$$

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = 6105,9.$$

$$\Rightarrow F = \frac{SS_R / 1}{SS_E / (n-2)} = 615,08$$

P - giá trị của phân bố F(1, 23) ứng với giá trị 615,08 bằng 0,000 nên ta cũng kết luận $b \neq 0$.

5.1.7. Tuyến tính hóa một số mô hình

Dùng phép biến đổi loga với biến hồi quy hay biến phản hồi, hoặc với cả hai, dùng phép nghịch đảo với biến hồi quy ..., ta có thể đưa một số mô hình về dạng tuyến tính.

Hồi quy logarith	$y = a + b.\ln x$
Hồi quy mũ	$y = a.e^{b.x}$ $(\Leftrightarrow \ln y = \ln a + b \ln x)$
Hồi quy lũy thừa	$y = a.x^{b}$ $(\Leftrightarrow \ln y = \ln a + b \ln x)$
Hồi quy nghịch đảo	$y = a + b.(1/x)$
Hồi quy tam thức	$y = a + bx + cx^2$

Chẳng hạn, khi cần dùng hồi quy mũ, trong phần chọn mô hình ta ấn $\boxed{\text{Exp}(3)}$; mọi thao tác khác tương tự.

• ***Sử dụng máy tính bỏ túi.*** Chúng ta mô tả ngắn gọn cách sử dụng máy tính bỏ túi CASIO fx-500MS để tính toán hồi qui. Dầu rằng những kết quả còn sơ lược so với các phần mềm chuyên dụng, song chúng cũng giúp ta nhất định trong công việc.

Xoá nhớ thống kê $\boxed{\text{SHIFT}}$ $\boxed{\text{MODE}}$ $\boxed{1}$ $\boxed{=}$

Gọi chương trình tính $\boxed{\text{MODE}}$ $\boxed{\text{REG}[3]}$

Chọn mô hình $\boxed{\text{Lin}[1]}$

Nhập dữ liệu. Chẳng hạn, cần nhập dữ liệu ở Ví dụ 5.1 ta ấn

$\boxed{45}$ $\boxed{,}$ $\boxed{24.2}$ $\boxed{\text{M}+}$

Cứ thế ta nhập cho hết dữ liệu.

Gọi kết quả. Nhập dữ liệu xong thì gọi kết quả. Việc gọi kết quả với biến x hoặc y: $\sum x_i^2, \sum x_i, \bar{x}, s_X, \tilde{s}_X, \sum y_i^2, \sum y_i, \bar{y}, s_Y, \tilde{s}_Y$ vẫn tiến hành như với thống kê 1 biến đã nêu ở cuối mục 4.1.2. Bảng 5.4 đưa ra vài tính toán như vậy cũng như một số tính toán khác.

Bảng 5.4. Một số thao tác phân tích hồi quy trên máy tính bỏ túi

Lượng cần tính	**Ấn**	**Kết quả**

$\sum x_i y_i$	SHIFT, S – SUM, ▷, $\sum$xy [3], =	9,265
s_Y	SHIFT, S – VAR, ▷, yσn [2], =	1,762
$\tilde{s}_Y$	SHIFT, S – VAR, ▷, yσn – 1 [3], =	1,904
$\hat{a}$	SHIFT, S – VAR, ▷, ▷, A [1], =	32,543
$\hat{b}$	SHIFT, S – VAR, ▷, ▷, B [2], =	-0.170
r_{XY}	SHIFT, S – VAR, ▷, ▷, r [3], =	-0.964
$\hat{x}(20)$	20, SHIFT, S – VAR, ▷, ▷, ▷, $\hat{x}$ [1]	73.78
$\hat{y}(70)$	70, SHIFT, S – VAR, ▷, ▷, ▷, $\hat{y}$ [2]	20.64

Sau khi có giá trị r_{XY}, dùng (5.1.31) ta tính được ƯL cho sai số chung $\hat{\sigma}^2$; tiếp theo ta có thể tính được T_a, T_b, ...

§ 5.2. MÔ HÌNH HỒI QUY TUYẾN TÍNH BỘI (1 tiết)

MHHQ tuyến tính bội là sự mở rộng tự nhiên của MHHQ tuyến tính đơn. Chúng ta ghi ra dưới đây những kết quả tóm tắt.

5.2.1. Phương trình hồi quy

a) Dạng quan sát và dạng ma trận

Giả sử mối quan hệ giữa biến phụ thuộc (biến phản hồi) Y và k biến độc lập (biến hồi quy) $x_1,...,x_k$ cho bởi mô hình

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon \qquad (5.2.1)$$

trong đó $\beta_0, \beta_1,...,\beta_k$ là các tham số chưa biết, gọi là các hệ số hồi quy, β_0 gọi là hệ số chặn, $\beta_1,...,\beta_k$ là các hệ số góc; ε là sai số ngẫu nhiên có kỳ vọng 0 và phương sai σ^2.

Khi không sợ nhầm lẫn, ta viết ngắn gọn (5.2.1) dưới dạng

$$E[Y \mid x_1,...,x_k] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (5.2.2)$$

hay đơn giản hơn nữa

$$E[Y] = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \qquad (5.2.3)$$

Để tìm hiểu mô hình (5.2.1) chúng ta tiến hành n quan sát và ghi lại kết quả dưới dạng bảng như Bảng 5.5.

Bảng 5.5. Số liệu cho mô hình hồi quy bội

y	x_1	x_2	.	x_k
y_1 .	x_{11} .	x_{12} .	. . .	x_{1k}

y_n	x_{n1}	x_{n2}		x_{nk}

Như vậy, dưới dạng quan sát, mô hình (5.2.1) viết lại dưới dạng

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + ... + \beta_k x_{1k} + \varepsilon_1 \\ . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ y_n = \beta_0 + \beta_1 x_{n1} + ... + \beta_k x_{nk} + \varepsilon_n \end{cases} \qquad (5.2.4)$$

Để thuận lợi cho ký hiệu và các phân tích tiếp theo, chúng ta sử dụng các ký hiệu ma trận sau đây.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ y_n \end{bmatrix}; \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ . \\ \beta_k \end{bmatrix}; \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ \varepsilon_n \end{bmatrix}$$

Khi đó, phương trình (5.2.4) được viết lại dưới dạng ma trận

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad (5.2.5)$$

trong đó **y** là n - véc tơ quan sát, **X** là ma trận cấp $n \times p$ của các biến độc lập $(p = k + 1)$ - còn gọi là ma trận kế hoạch - **β** là p - véc tơ các hệ số hồi quy, **ε** là n - véc tơ sai số ngẫu nhiên.

b) Tuyến tính hóa một số mô hình

Mô hình (5.2.3) là tuyến tính vì nó tuyến tính với các tham số β_i. Trong ứng dụng chúng ta thường gặp mô hình dạng

$$E[Y] = \beta_1 g_1(x_1,...,x_\ell) + ... + \beta_p g_p(x_1,...,x_\ell) \quad (5.2.6)$$

trong đó $g_1,...,g_p$ là các hàm nào đó của các biến hồi quy $x_1,...,x_\ell$.

Đây là mô hình tuyến tính với các tham số β_i, phi tuyến với các biến $x_1,...,x_\ell$. Xét phép đổi biến

$$z_1 = g_1(x_1,...,x_\ell);...;\ z_p = g_p(x_1,...,x_\ell).$$

Ta có thể đưa (5.2.5) về dạng thông thường

$$E[Y] = \beta_1 z_1 + ... + \beta_p z_p \qquad (5.2.7)$$

là mô hình tuyến tính với cả tham số lẫn các biến hồi quy. Như vậy từ nay ta vẫn gọi mô hình (5.2.6) là tuyến tính. Xét một số trường hợp đặc biệt.

b1. Hồi quy đa thức. Xét mô hình

$$E[Y] = a_0 + a_1 x + ... + a_k x^k.$$

Đặt $z_1 = x; ...;\ z_k = x^k$, ta đưa mô hình này về dạng

$$E[Y] = a_0 + a_1 z_1 + ... + a_k z_k.$$

Đặc biệt, người ta hay xét mô hình tam thức và đa thức bậc ba:

$$E[Y] = a + cx + cx^2,$$

$$E[Y] = a + cx + cx^2 + dx^3.$$

b2. Mô hình đa thức bậc 2 của hai biến. Đó là mô hình

$$E[Z] = a + bx + cy + dx^2 + exy + fy^2.$$

Đây là mô hình tuyến tính với 6 tham số a, b, c, d, e, f. Trường hợp giả thuyết $e = 0$ bị bác bỏ, ta nói hai biến hồi quy x và y là tương tác với nhau, mô hình có chứa số hạng tích chéo xy. Trái lại, nếu $e = 0$, ta nói mô hình không chứa số hạng tích chéo xy, 2 biến x và y là không tương tác với nhau.

b3. Dùng phép biến đổi loga với biến phản hồi

b4. Hồi quy có chứa sin, cos.

Giả sử biến phụ thuộc có dạng

$$Y(t) = a + bt + c\sin t + d\cos t + \varepsilon.$$

Bằng cách đặt $x_1 = t;\ x_2 = \sin t;\ x_3 = \cos t$, ta đưa mô hình về dạng tuyến tính thông thường.

5.2.4. Ước lượng và dự đoán

a) Khoảng tin cậy cho tham số đơn lẻ

b) Khoảng tin cậy cho đáp ứng trung bình.

c) Dự đoán cho quan sát mới. ƯL điểm của dự đoán cho quan sát tương lai tại mức $x_{01}, ..., x_{0k}$ của các biến độc lập là

$$\hat{y}_0 = \mathbf{x}_0^T \beta = \beta_0 + \beta_1 x_{01} + ... + \beta_k x_{0k}.$$

Khoảng dự đoán $100(1-\alpha)\%$ cho quan sát tương lai này là

$$\hat{y}_0 \pm t_{\alpha/2}(n-p)\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}. \quad (5.2.17)$$

5.2.6. Sử dụng phần mềm

Các phần mềm thống kê ngày nay cho phép phân tích mô hình với số biến hồi quy lên đến hàng ngàn và số quan sát lên đến hàng chục vạn. Chúng ta cần có những kiến thức cơ bản để tận dụng những lợi thế của các phần mềm này. Mỗi phần mềm có những thế mạnh của nó, song chúng đều có phần phân tích hệ số và phân tích phương sai. Chúng ta tìm hiểu sơ bộ qua một vài ví dụ.

Ví dụ 5.3 *(Phân tích số liệu lực kéo).* Chúng ta lấy lại ví dụ lực kéo ở Ví dụ 5.2. Giả sử chúng ta đã nhập số liệu vào cửa sổ biên tập dữ liệu. Sau đây là một số thao tác cơ bản.

Bảng 5.5. Kết quả xử lý với số liệu lực kéo dây dẫn

TT	Lực kéo y_i	Độ dài x_1	Độ cao x_2	Dự báo $\hat{y}_i$	Phần dư e_i	Phần dư chuẩn hóa d_i
1	9.95	2	50	8.38	1.57	.687
2	24.45	8	110	25.60	-1.15	-.501
3	31.75	11	120	33.95	-2.20	-.963
4	35.00	10	550	36.60	-1.60	-.698
5	25.02	8	295	27.91	-2.89	-1.265
6	16.86	4	200	15.75	1.11	.487
7	14.38	2	375	12.45	1.93	.843
8	9.60	2	52	8.40	1.20	.523
9	24.35	9	100	28.21	-3.86	-1.689
10	27.50	8	300	27.98	-.48	-.208
11	17.08	4	412	18.40	-1.32	-.578
12	37.00	11	400	37.46	-.46	-.202
13	41.95	12	500	41.46	.49	.215
14	11.66	2	360	12.26	-.60	-.263
15	21.65	4	205	15.81	5.84	2.553
16	17.89	4	400	18.25	-.36	-.158
17	69.00	20	600	64.67	4.33	1.894
18	10.30	1	585	12.34	-2.04	-.890
19	34.93	10	540	36.47	-1.54	-.674
20	46.59	15	250	46.56	.03	.013
21	44.88	15	290	47.06	-2.18	-.953
22	54.12	16	510	52.56	1.56	.681
23	56.63	17	590	56.31	.32	.141
24	22.13	6	100	19.98	2.15	.939
25	21.15	5	400	21.00	.15	.067

Ta thấy hệ số xác định $R^2 = 0,981$, vậy có 98,1% số liệu được giải thích bởi mô hình; đây là một tỷ lệ khá lớn. ƯL cho phương sai chung của mô hình là $\hat{\sigma}^2 = 2,2881^2$. Mức ý nghĩa của thống kê F là 0,000, rất nhỏ so với 0,01: Mô hình có tác dụng tốt để giải thích số liệu. Tất cả các mức ý nghĩa của thống kê T của các tham số đều nhỏ hơn 0,05 (giá trị cực đại 0,044 ứng với biến hằng số). Hậu quả là khoảng tin cậy của tất cả các hệ số đều không chứa gốc tọa độ. Như vậy, các kiểm định T không bác bỏ mô hình. Mô hình dự tuyến là

$$Y = 2,264 + 2,744x_1 + 0,013x_2 + \varepsilon \qquad (*)$$

BÀI TẬP: Kiểm định phi tham số (2 tiết)

b) Thảo luận	
c) Tự học	
d) Bài tập chuẩn bị tối thiểu	
Tài liệu	Tài liệu [1], tr ....

Bài giảng 15: Mô hình hồi quy tuyến tính (tiếp)

Chương, mục: 5

Tiết thứ: 57-60 Tuần thứ: 15

Mục đích, yêu cầu:

- Củng cố kiến thức
- Sẵn sàng để thi cuối học kỳ

- Hình thức tổ chức dạy học:

Hình thức chủ yếu: Lý thuyết, thảo luận - tự học, tự nghiên cứu

- Thời gian:

Lý thuyết, thảo luận: 4t - Tự học, tự nghiên cứu: 4t

- Địa điểm:

Giảng đường do P2 phân công.

- Nội dung chính:

- Bài tập Hệ số tương quan (1tiết)
- Mô hình hồi quy tuyến tính (1 tiết)
- Chữa các bài chưa có điều kiện chữa (2tiết)
  (Giáo viên làm là chính)
  Ôn tập chuẩn bị thi hết môn
- Nhắc lại về các câu hỏi lý thuyết, cách học chúng.
- Một số kinh nghiệm khi thi
- Nhắc lại tinh thần nghiêm túc trong thi cử
- Nhắc một số quy đinh trong kỳ thi

Ôn tập